Evaluación de métodos de agrupamiento DBSCAN y LDA para el análisis de contenido de la red social Twitter.

Diana Yamile Olarte Sierra y Nicolás Ariza Pardo

Trabajo de grado para optar al título de ingeniero industrial

Director
Daniel Orlando Martínez Quezada
M.sc

Codirector
Henry Lamos Díaz
PhD

Universidad Industrial de Santander
Facultad de Ingenierías Físicomecánicas
Escuela de Estudios Industriales y Empresariales
Bucaramanga
2020

AGRADECIMIENTOS

A nuestro director Daniel Orlando Martínez Quezada por la orientación, confianza y dedicación durante el desarrollo de nuestro proyecto.

A nuestro codirector Henry Lamos Díaz por el apoyo y tiempo dedicado durante el desarrollo del proyecto.

A Raúl Valerio por su disposición y orientación profesional.

Al grupo OPALO por brindarnos la oportunidad de incursionar en la investigación, dotándonos de herramientas necesarias para llevar feliz término nuestros objetivos.

Tabla de contenido

Lista de tablas

Lista de figuras

Apéndice

Apéndice A. Artículo de Investigación

RESUMEN

Título del proyecto: evaluación de métodos de agrupamiento DBSCAN y LDA para el análisis de contenido de la red social Twitter*.

Autores: Diana Yamile Olarte Sierra, Nicolás Ariza Pardo**.

Palabras clave: aprendizaje automático, Twitter, minería de texto, agrupamiento, LDA, DBSCAN.

Descripción:

En la actualidad debido al gran volumen de datos que se manejan es importante usar métodos de monitoreo no supervisado que permitan encontrar relaciones entre diferentes temas y se obtenga información valiosa sobre las discusiones que son tendencia. Dicha información tiene aplicaciones en los diferentes sectores de la industria siendo vital en la toma de decisiones estratégicas. En el presente trabajo se hace un análisis a una base de datos de prueba constituida a través de la red social Twitter, estos datos fueron recolectados a lo largo de cinco meses a través de la página oficial en Twitter de la Universidad Industrial de Santander. Se utiliza la herramienta Rstudio del lenguaje de programación R con la cual se realiza el preprocesamiento de los datos y se define el corpus, además se usa el método del codo y se aplican las métricas Griffits2004, CaoJuan2009, Arun2010 y Deveaud2014 para la definición de los parámetros iniciales para la ejecución de los algoritmos usados. Con esto se hace una evaluación de dos de los métodos de agrupamiento, LDA y DBSCAN aplicando el coeficiente de silueta para analizar la calidad y el grado de coherencia de los grupos obtenidos e identificar tendencias y temas relevantes en la información extraída.

* Trabajo de grado.

** Facultad de Ingenierías Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: M.sc. Daniel Orlando Martínez Quezada. Codirector: PhD. Henry Lamos Díaz.

ABSTRACT

Project title: evaluation of DBSCAN and LDA clustering methods for content analysis of the social network Twitter[*].

Authors: Diana Yamile Olarte Sierra, Nicolás Ariza Pardo[**].

Keywords: machine learning, Twitter, data mining, cluster, LDA, DBSCAN.

Description:

At present, due to the large volume of data that is handled, it is important to use unsupervised monitoring methods that allow finding relationships between different topics and obtain valuable information on the discussions that are trending. This information has applications in different sectors of the industry, being vital in making strategic decisions. In this work, an analysis is made of a test database constituted through the social network Twitter, these data were collected over five months through the official Twitter page of the Industrial University of Santander. The Rstudio tool of the R programming language is used with which the data is preprocessed and the corpus is defined, in addition the elbow method is used and the metrics Griffits2004, CaoJuan2009, Arun2010 and Deveaud2014 are applied to define the initial parameters for the execution of the algorithms used. With this, an evaluation of two of the grouping methods, LDA and DBSCAN, is made, applying the silhouette coefficient to analyze the quality and degree of coherence of the groups obtained and identify relevant trends and themes in the information extracted.

[*] Degree Project.

[**] Faculty of Physico mechanical Engineering. Industrial and Business School. Director: M.sc. Daniel Orlando Martínez Quezada, Codirector: PhD. Henry Lamos Díaz

Introducción

En la actualidad el acceso a internet ha permitido que las personas puedan utilizar datos de manera inmediata y en tiempo real, esto es gracias al amplio portafolio de recursos y servicios de información que ofrece la red, que está organizada en documentos de hipertexto entrelazados y aplicaciones de la World Wide Web (WWW), redes sociales, correo electrónico, la transmisión de archivos, la transmisión de contenido y comunicación multimedia, el acceso remoto, los juego en línea, etc. Dichos recursos generan a diario datos que al ser procesados de manera adecuada podrían aportar conocimientos útiles sobre el comportamiento de las personas que tienen la posibilidad de utilizar la red, es en este punto donde la minería de datos adquiere un papel muy importante.

La minería de datos es la búsqueda de patrones que pueden existir en grandes bases de datos, pues en estas masas de datos se encuentra información valiosa, sorprendente, novedosa e inesperada. Para realizar un análisis de datos se pueden utilizar técnicas de agrupamiento, que se pueden ver como una forma de modelado de datos que proporcionan un resumen de estos conjuntos de datos. Por lo tanto, el agrupamiento de datos tiene relación con muchas disciplinas y tiene una amplia gama de aplicaciones. (Berkhin, 2006)

Dado el gran número de datos que fluye a través de las redes sociales actualmente, se hace necesario el estudio de la efectividad de los algoritmos usados para el análisis de dicha información, teniendo en cuenta que los algoritmos que se van a implementar en el proyecto (LDA y DBSCAN). Para el algoritmo de agrupamiento DBSCAN (agrupación espacial basada en la densidad de aplicaciones con ruido) que se ha usado ampliamente durante los últimos años en muchas áreas debido a su simplicidad y su ventaja de encontrar agrupaciones de diferentes tamaños

y formas, se vuelve inestable cuando se detectan objetos de borde de agrupaciones adyacentes (Tran, Drab, & Daszykowski, 2013). Por lo tanto, lo que se busca es comparar los resultados obtenidos para garantizar la calidad de la información obtenida. (Sepúlveda, 2016)

El presente proyecto tiene como fin la comparación entre dos modelos de agrupamiento, por lo que se quiere evaluar el grado de cohesión y separación de los modelos de agrupación particional LDA y DBSCAN, y de esta forma determinar cuál de los dos modelos presenta mejores resultados, es decir cual minimiza la relación entre grupos y cual maximiza la relación entre los miembros de un grupo.

Con el desarrollo de este proyecto se busca establecer un procedimiento para la evaluación de modelos de agrupamiento particional (LDA y DBSCAN) los cuales podrían tener una participación importante en el análisis de datos para el estudio de percepción de marca y sentimientos respecto a diferentes industrias que tienen los usuarios de la red social Twitter, y a partir de esto se lograría plantear estrategias de marketing que mejoren la imagen de cara al cliente que tienen las compañías que estén interesadas en el análisis de contenido (Xia Liu; Alvin C. Burns; Yingjian Hou, 2017). Otra área de aplicación de estos modelos es la gestión de la cadena de suministros, pues con dichos datos se respalda la toma de decisiones para que la cadena de suministro sea más dinámica, ágil y global (Canzanello; del Canal; Rossmann, 2018). Además, un buen procesamiento de contenido de redes sociales lograría generar información valiosa para la investigación y la práctica en gestión de operaciones, un claro ejemplo es el diseño de producto o de procesos (HK, Chan; E, Lacka; R, Yee; MK, Lim, 2017).

Tabla 1.

Cumplimiento de los objetivos del proyecto

Objetivos Específicos	Cumplimiento
• Realizar una revisión bibliográfica de las técnicas de agrupamiento de texto y su uso para el análisis de contenido de la red social Twitter.	Capítulo 4
• Consolidar una base de datos a partir de la red social Twitter para el desarrollo del caso de estudio.	Capítulo 6
• Evaluar los modelos de agrupamiento LDA y DBSCAN en el conjunto de datos consolidados para analizar el contenido de la plataforma Twitter.	Capítulo 7
• Elaborar un artículo de carácter publicable en base a los resultados obtenidos durante la investigación.	Apéndice A

1. Planteamiento del problema

La cuarta revolución industrial trae consigo una serie de cambios que llevarán a las empresas a replantear sus modelos de negocio y adaptarlos a la era digital que se está viviendo. El surgimiento de conceptos como la inteligencia artificial, el big data y el internet de las cosas facilita la obtención de información de fuentes primarias (usuarios de las redes sociales para este estudio) dando la

posibilidad a las organizaciones de conocer el mercado y evaluar si sus estrategias responden a las necesidades de sus clientes.

Twitter es una de las redes sociales más populares en la actualidad, brinda a sus usuarios un espacio para compartir su opinión acerca de temas de interés o situaciones coyunturales. Dada su naturaleza pública es utilizado como medio de propagación y transmisión de datos, lo que convierte a esta plataforma digital en una de las fuentes públicas de información en tiempo real más grandes. Además, su fácil acceso a sus bases de datos mediante su interfaz de programación de aplicaciones (API) permite su transferencia a otro software para su tratamiento y análisis.

La minería de datos a través de métodos como el aprendizaje automatizado y la inteligencia artificial permite el análisis de grandes volúmenes de datos. Existen algoritmos para el agrupamiento y análisis de contenido cuyo objetivo principal es la identificación de patrones y tendencias, para este caso de estudio se propone la aplicación de los métodos de agrupamiento LDA y DBSCAN evaluando su efectividad y la calidad de la información obtenida.

Ésta información es muy valiosa para las empresa pues tiene un gran número de aplicaciones en las diferentes industrias, pues permite conocer la percepción de marca que tienen los clientes acerca de una organización en particular, verifica si las condiciones están dadas para la inversión en ciertos lugares, identifica qué problemáticas pueden afectar las cadenas de suministro, provee una base de apoyo para la toma de decisiones estratégicas, evalúa la viabilidad de la creación de un nuevo producto, etc. Brindando una ventaja competitiva que hace que las organizaciones agreguen valor a sus procesos.

2. Justificación del proyecto

El excesivo crecimiento de los datos que se maneja en la actualidad en los sistemas productivos ha sobrepasado la capacidad humana de analizar, resumir y extraer información de grandes bases de datos que puedan brindar alternativas de mejora y ventajas competitivas a una organización. El aprendizaje no supervisado se convierte en una herramienta estratégica a la hora de tomar decisiones en todos los niveles de la organización (mercadeo, producción, administración, logística, etc.). En la actualidad las empresas colombianas poseen un gran volumen de datos que se encuentran almacenados y guardados de forma incorrecta, lo que ocasiona que la información que estos contengan no se aproveche de manera adecuada (Echeverri, y otros, 2014).

Existen algoritmos de agrupamiento supervisados y no supervisados que brindan a las empresas una fuente eficiente de tratamiento y procesamiento de información a partir de grandes volúmenes de datos. Muchos de estos datos se encuentran actualmente en las plataformas digitales, las cuales poseen un gran tráfico de usuarios que opinan sobre los temas que son tendencia y otorgan elementos que con un adecuado análisis proveen de conocimiento útil en los procesos industriales.

En el presente documento se evalúa el desempeño de dos algoritmos de agrupamiento en el tratamiento de una base datos a través de la API de Twitter, comparando la coherencia de los resultados obtenidos y su comportamiento en el caso de estudio.

3. Objetivos

3.1 Objetivo general

Comparar los resultados de los algoritmos de agrupamiento LDA y DBSCAN para el análisis de contenido a partir de datos de la plataforma Twitter.

3.2 Objetivos específicos

- Realizar una revisión bibliográfica de las técnicas de agrupamiento de texto y su uso para el análisis de contenido de la red social Twitter.
- Consolidar una base de datos a partir de la red social Twitter para el desarrollo del caso de estudio.
- Evaluar los modelos de agrupamiento LDA y DBSCAN en el conjunto de datos consolidados para analizar el contenido de la plataforma Twitter.
- Elaborar un artículo de carácter publicable en base a los resultados obtenidos durante la investigación.

4. Revisión de literatura

La mineria de datos y más especificamente el análisis de cotenido ha tenido un gran número de aplicaciones y usos en diferentes áreas de investigación. El aprendizaje automático ha sido usado para clasificar datos principalmente en la red social twitter identificando y catalogando tweets en temas relevantes con el objetivo de analizar el grado en que los usuarios liberan sus emociones,

reflejando una imagen de su comportamiento a través de las palabras que usan en sus publicaciones. (Zhang, Dong, & Mu, 2018)

La extracción de temas latentes de conjuntos de textos cortos (tweets) es una tarea muy importante para la aplicación de información basado en temas de contenido, dentro de esta información se encuentra la caracterización del usuario, detección del tema y el interés del usuario. Para dicho modelado de temas se parte de la utilización de técnicas como LDA y PLSA que suponen encontrar correlación entre múltiples palabras que tienen relación. Aunque conocer la relación entre textos cortos es difícil, encontrar un punto en común dentro de un grupo de datos puede generar temas precisos y resultados de alta calidad (Rashid, Muhammad, Shah, & Irtaza, 2019).

Con el auge de las redes sociales durante los últimos años estas se han covertido en un asunto de estudio importante para muchas disciplinas (biología, economía, psicología y aprendizaje auntomático), el interés que despiertan las redes sociales recide en su variedad de aplicaciones, pues gracias a su abundancia de datos (conocimiento) se puede evaluar cuáles son los temas de interés y los de mayor influencia, la explicación de los diferentes tipos de enfoques basados en la centralidad dan la opción de expresar una explicación sociológica de los datos y técnicas de influencia y propagación de la información (Coumo, Salvatore; Maiorano, Francesco, 2018).

La detección de comunidad en redes sociales permite identicar las tendencias de los usuarios cuando interactúan en plataformas digitales, las diferentes disciplinas han usado diferentes métodos para caracterizar el enfoque comunitario del marketing viral que se basa en los intereses de los sujetos que usan las redes sociales para compartir intereses. Además, si se estudia los intereses comunes entre el usuario y su grupo de amigos se puede obtener conocimiento que ayuda

a las diferentes áreas a plantear sus estrategias en base a los gustos de su población objetivo (Alsuwaidan & Ykhlef, 2018).

Las redes sociales son estructuras dinámicas que están sujetas a cambios continuamente, el concepto de comunidad varía en el tiempo, ya que, los usuarios usan o dejan de usar las plataformas digitales de manera abrupta. Con la ayuda de la técnica de resonancia adaptativa se hace un estudio sobre la comunidad y se trata de explicar a través de este algoritmo dinámico como afectan estos cambios a la red con alta precisión. El modelado ARTISON (detección de comunidad inspirada en ART en la red social), es un algoritmo basado en la represantación dinámica de la red neuronal ART, que se desarrolla utilizando capacidades de reconocimiento humano flexibles y cognitivos. La aplicación de este algoritmo arroja como resultado la detección de cambios bajos y abruptos en la red al tiempo que arroja la cantidad de comunidades (Sadat & Ali, 2017).

Las redes sociales funcionan como un filtro para la información publicada, puesto que, los usuarios deciden que publicaciones quieren ver y compartir. Campañas de marketing, que son difundidas especialmente a través de redes sociales, tienen un mayor impacto dado que en varias ocasiones llega a un público mucho más amplio que al grupo focal inicial. Esto ha puesto las bases para que las empresas en el futuro difundas las campañas relacionadas con lazamientos de nuevos productos y servicios a través de medios virtuales principalmente, garantizando mejores resultados (Chu, y otros, 2015).

Los usuarios de las redes sociales comparten en sus perfiles información (carrera, universidad, pasatiempos…) que puede ser usada para agruparlos y dar recomendaciones de acuerdo a los datos suministrados por los individuos. Inferir en los atributos de los usuarios de las paltaformas digitales ha sido un interesante tema de estudio para el aprendizaje supervisado, la idea de tener acceso a la información de los usuarios es encontrar patrones ocultos en los usuarios y agruparlos de acuerdo

a dichos patrones (caracterizar los usuarios y agruparlos de acuerdo a intereses comunes) (Ding, Yan, Zhang, Dai, & Dong, 2016).

El análisis de contenido ha arrojado buenos resultados en estudios relacionados con el comportamiento de las personas en situaciones cotidianas o problemáticas coyunturales muy especificas como la primavera árabe. Combinando el análisis de contenido con las técnicas automatizadas de análisis de red para determinar los roles desempeñados por usuarios de Twitter para comunicarse especificamente durante el levantamiento del Parque Gezi en Turquía permitiendo detectar la posición y participación de dichos usuarios dentro de la protesta (Ogan & Varol, 2017).

Mediante el estudio de comportamiento de usuarios en redes sociales tambien se ha podido predecir y detectar criminales potenciales como lo presentan en su estudio los autores Cesur, Ramazan y otros, que por medio del uso de las tecnologias de aprendizaje automático y análisis de big data lograron analizar e identificar palabras claves dentro de las publicaciones de usuarios arrojando una tasa de éxito alrededor de 71.61% en la detección de criminales cibernéticos principalmente (Cesur, Ceyhan, Kermen, & Sagiroglu, 2017).

En cuanto al área de ingeniería industrial estudios como el realizado por Araujo, Theo; Kollat, Jana enfocado principalmente a industrias alimentarias permitio conocer la importancia de la difusión de información relacionada con la responsabilidad social corporativa a traves de twitter demostrando que las organizaciones con mayor presencia en dichas plataformas tienen un mayor grado de difusión y aceptación entre su mercado (Araujo & Kollat, 2018).

Gracias a la aparición de sistemas expertos cada vez más sofisticados se ha conseguido, a lo largo de los últimos años, que la capacidad de generar datos y recolectarlos sea mucho más grande.

El uso de la minería de datos (data mining) permite descubrir conocimiento en grandes volúmenes de datos, dicho conocimiento es usado por las organizaciones a la hora de tomar decisiones estratégicas y tácticas a partir de la automatización de los sistemas (Marcos, 2004).

(Neme-chaves, 2018) En su investigación titulada: "Minería de texto a través de Twitter de la marca Juan Valdéz café" de la universidad Santo Tomás, realizó una recolección de datos a través de The R Project for Statistical Computing, donde a partir de 243 tweets evaluó las palabras que asociaban a dicha marca, las palabras más frecuentes en el corpus fueron: "café", "colombiano", "saludable", "Valdéz" y "campeón". A partir de los datos recolectados, se dividió el corpus de análisis en dos grupos usando el método de asignación latente de Dirichlet (LDA), pues este modelado le permitió al autor estimar tanto los temas como las palabras al mismo tiempo y así encontrar su relación. Los tweets recolectados mostraron que respecto al café Juan Valdéz existían temas relacionados al producto, el sabor, la naturaleza, la siembra, entre otros. Dando gran relevancia al disfrute y la libertad que les daba el producto a los consumidores.

La minería de datos se ha convertido en una poderosa herramienta para la toma de decisiones de mercadeo, producción, organización y otros factores que hacen a una compañía más competitiva. La minería de datos hace parte importante de la investigación de operaciones, sin embargo esta puede incidir en el estudio de estrategias de mercadeo B2B (Buiness to Business). Una de las vemtajas de la minería de datos es su facilidad de uso, para lo cual se hace necesario la implementación de los algoritmos adecuados (dependen de la necesidad), la evaluación adecuado de los datos aumenta la utilidad de la información obtenida después de un procesamiento de la información. Para ello se debe obtener indicadores sobre cuatro facetas del resultado: bondad de ajuste, relevancia, novedad y aplicabilidad. Los consumidores están cada vez más informados sobre los diferentes productos y servicios que se ofrecen en el mercado, por lo tanto, a partir de la

minería de datos es posible plantear estrategias que permitan cubrir la necesidad del consumidor e incremente la utilidad de las empresas, así mismo reducir costos y predecir eventos futuros que podrían impactar negivamente a la organización (Altamiranda et al., 2013).

5. Marco teórico

En este capítulo se describirán los principales conceptos para tener en cuenta para el desarrollo de este proyecto de investigación.

5.1 Inteligencia artificial

"Una definición comúnmente aceptada relaciona la disciplina de la Inteligencia Artificial (IA) con el análisis y el diseño de sistemas artificiales autónomos capaces de exhibir un comportamiento inteligente. Se asume que, para que un agente actúe inteligentemente, debe poder percibir su entorno, elegir y planificar sus objetivos, actuar hacia la consecución de estos objetivos aplicando algún principio de racionalidad e interactuar con otros agentes inteligentes, sean estos artificiales o humanos" (Mar, n.d.; Rygielski, Wang, & Yen, 2002).

La inteligencia artificial no es una disciplina nueva, ya desde mediados de los años cincuenta se acuñó dicha denominación. Fueron Marvin Minsky, John McCarty, Nathan Rochester y Claude Shannon los que convocaron a una convención en 1956 en la cual el principal eje de discusión giraba en torno a la premisa de que todos los aspectos del aprendizaje o cualquier característica de la inteligencia se podrían construir a través de una máquina que podría simularlos. Fue a partir de

esta reunión que surgió el término de inteligencia artificial del cual al día de hoy se han realizado numerosos avances (Mar, n.d.; Rygielski et al., 2002).

5.2 Minería de datos

La minería de datos es una disciplina que toma como una interacción de la estadística, tecnología de bases de datos, reconocimiento de patrones, aprendizaje automático y otras áreas, que se ocupa del análisis secundario de grandes bases de datos con el fin de encontrar relaciones, que por su gran tamaño no son fáciles de observar, de interés o valor para las personas que son dueñas de las bases de datos (Rygielski et al., 2002).

5.2.1. Problemas que aborda la minería de datos

La minería de datos aborda cualquier problema susceptible a ser tratado mediante técnicas de data mining (Aluja, 2001). Entre ellos se encuentran:

Búsqueda de lo inesperado por descripción de la realidad multivariante. Entre más variables se tengan más ricas, precisas, coherentes y globales serán las descripciones y será más fácil predecir o detectar lo inesperado.

***Búsqueda de asociaciones*.** Que tan está relacionado un hecho con otro, si dependen entre ellos o es fácil de recomendar un producto si se ha adquirido otro.

Definición de tipologías. Los consumidores son infinitos, pero se pueden agrupar en familias dependiendo de sus perfiles y la población a la que pertenecen. Este tipo de tipologías pueden ser de consumo, opiniones, valores, estilos de vida, etc.

Detección de ciclos temporales. Identificar el ciclo de necesidades que llevan a realizar compras a lo largo de su vida.

Predicción. ¿Cuál es la probabilidad de que ocurra un hecho?

5.2.2. Proceso de minería de datos

Es necesario conocer cuáles partes del proceso son las que se pueden automatizar y cuáles no. La identificación de los datos que se desean tratar es el punto de partida de la minería de datos, es clave saber cuáles datos se necesitan, dónde se pueden encontrar, cómo obtenerlos y para qué se pueden usar. Este filtro depura los datos, deja aquellos que realmente tienen impacto en la investigación que se quiere realizar y desecha aquellos que hacen ruido en el estudio.

Después de seleccionados los datos, se procede a un pre-procesamiento donde se eliminan datos incorrectos, no válidos y que son desconocidos (o que no se adaptan al algoritmo de agrupamiento que se quiera implementar). Debido a la gran cantidad de datos que se manejan en la actualidad se pueden tener inconsistencias, por lo que, la baja calidad de los datos trae consigo resultados de mala calidad (Han, Kamber, & Pei, 2011). El pre-procesamiento es uno de los componentes clave en el estudio de la minería de datos, pues es aquí donde se limpian y se estructuran los datos para su posterior análisis (Feinerer, Hornik, & Meyer, 2008). La selección de características de las bases de datos, permite que se elijan las variables más influyentes en el problema, garantizando el no sacrificio de la calidad del modelo de conocimiento aplicado.

Los métodos para la selección de características son básicamente dos:

- Aquellos basados en la elección de los mejores atributos del problema.
- Aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

Una vez planteado esto, se procede a la extracción de conocimiento, mediante una técnica de minería de datos se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables.

Ya obtenido el modelo, el paso a seguir es validar si los resultados y las conclusiones que se encontraron son válidas y satisfactorias para el estudio que se realizó. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si los resultados no son los esperados es necesario generar nuevos modelos que adapten mejor a la situación que se presenta (Beltrán Martínez, 2014).

Lo anteriormente descrito se puede ver gráficamente en la figura 6, donde se muestra el proceso de minería de datos:

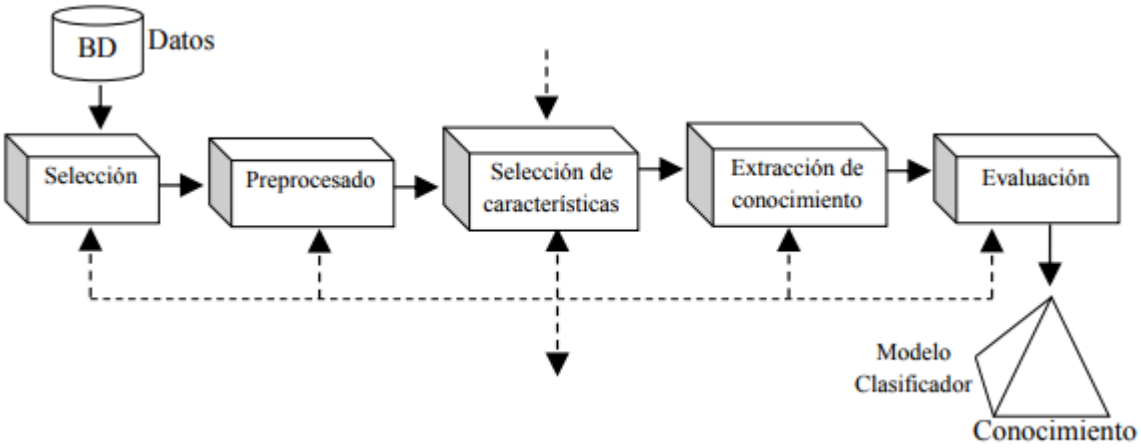

Figura 1. Proceso de minería de datos. Tomado de Minería de datos Beltrán Martínez, Beatriz (2014)

5.2.3. Técnicas de minería de datos.

Gracias a la minería de datos se ha logrado cambiar el enfoque que se tenía sobre el análisis de los mismos, pues se dejó de implementar simplemente como verificación y se está usando, en la actualidad, como una poderosa herramienta de descubrimiento de conocimiento. La hipótesis que se plantea para la validación de datos no es necesaria cuando se quiere descubrir conocimiento, pues sin necesidad de esta se puede extraer información valiosa a partir de un buen procesamiento de datos. La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos, razón por la cual esta técnica es mucho más eficiente que el análisis dirigido a la verificación cuando se intenta explorar datos procedentes de bases de alta complejidad y gran tamaño. Estas técnicas están en evolución continúa debido a la interacción de diferentes disciplinas o campos de investigación como lo son bases de datos, el reconocimiento de patrones, la inteligencia artificial, los sistemas expertos, la estadística, la visualización y la recuperación de información.

Los algoritmos de minería de datos están divididos en dos grupos: supervisados (predictivos) y los no supervisados (destinados al descubrimiento del conocimiento).

Algunas de las técnicas de minería de datos en ambas categorías son las siguientes:

Tabla 2.

Técnicas de minería de datos.

SUPERVISADOS	**NO SUPERVISADOS**
Árboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento (Clustering)
Series temporales	Reglas de posición
	Patrones secuenciales

Nota: adaptado de Beltrán Martínez, (2014)

5.3. Clustering

El clustering o segmentación es una herramienta que permite la identificación de grupos donde los elementos que lo componen tienen similitud entre ellos, pero que dichos elementos comparados con los de otros grupos guardan diferencias. Una forma de aprendizaje no supervisado es la agrupación (en inglés, clustering) entre los que se pueden señalar COBWEB, EM, K-Means. El clustering se encarga de partir los datos en grupos con características similares, con lo que se quiere medir las similitudes entre los elementos, para esto se hace uso de diferentes métricas de distancia, tales como: la distancia Euclídea, Manhattan, Máximum, Minkowski, etc. La agrupación de datos tiene amplia aplicación en la minería de datos, pues se puede realizar una exploración de datos científicos, recuperar información que se encuentra oculta en grandes volúmenes de datos, la minería de texto, el procesamiento de datos procedentes de bases de datos de páginas web, entre otros usos (Garre, M., Cuadrado, J., Sicilia, M. A., Rodríguez, D., & Rejas, R. 2007).

La idea de agrupar los datos es encontrar que los miembros del conglomerado tienen más parecido entre ellos que con los miembros de otros grupos. El agrupamiento trata de analizar las asociaciones que se encuentran, dichas asociaciones son de alta calidad si la semejanza en clusters es baja y la similitud entre los elementos del grupo es alta. También, se pueden identificar datos atípicos que no tienen ningún tipo de relación con los demás grupos.

5.3.1. Clustering de texto.

Con el clustering de texto se logra identificar la estructura dentro de una colección de documentos, logrando así identificar tendencias o temas que tienen más frecuencia en un conglomerado y calcificándolos en diferentes grupos. La finalidad (como en los métodos de cluster numéricos) es maximizar la similitud entre los elementos de un mismo clúster a la vez que maximiza las diferencias entre los otros conglomerados (Jing, L. 2008).

5.3.2. Análisis de clúster.

El análisis de clúster es una variedad de técnicas que se usan para buscar grupos en un conjunto de datos o elementos. Para la agrupación de datos se inicia a partir de un conjunto de datos que al parecer no tienen algún tipo de relación, pero que se pueden partir en diferentes grupos (clúster) que son resultado de particiones del grupo original. Las leyes matemáticas por las que se rigen estos métodos tienen el nombre de “taxonomía numérica”, definida como el conjunto de leyes formalizadas que intentan construir clasificaciones basadas en la semejanza observado entre los elementos a clasificar (Fernández, O. 1991).

5.3.2.1. Propiedades de los clústers

- *Densidad.* Define como un conglomerado espacial de puntos relativamente compacto en comparación con otras áreas del espacio que tienen menos o ningún punto.
- *Varianza.* Grado de dispersión de los puntos de cada conglomerado en el espacio.
- *Forma.* Configuración espacial de los puntos.
- *Separación.* Grado de solapamiento entre los clústers.

5.4. Algoritmos de agrupamiento

Para conocer y tipificar el comportamiento de una población se importante poder clasificar un grupo de datos en diferentes clústers cuyos elementos guarden una fuerte relación entre ellos, si se quiere lograr encontrar conocimiento sobre un grupo de datos es importante usar métodos de agrupamiento. Existen dos tipos de agrupamiento: agrupamientos paramétricos y agrupamientos no paramétricos (Pascual, Pla, & Sánchez, n.d.-a; Rygielski et al., 2002).

5.4.1. Algoritmos de agrupamiento paramétricos

5.4.1.1. Mixturas finitas. Es una herramienta para modelar probabilidad de densidades de un grupo de individuos (con una o varias variables) que se produjeron a partir de fuentes aleatorias y derivan los parámetros de estas fuentes, lo que permite identificar la procedencia de cada observación. Sea $Y = [Y_1, Y_2, \dots, Y_d]^T$una variable aleatoria d-dimensional con y = $[y1, y2, \dots, yd]^T$ representando un resultado particular de Y. Se dice que Y sigue una distribución de mixtura finita con k componentes si su función de densidad de probabilidad se puede escribir por:

$$p\left(\frac{y}{\theta}\right) = \sum_{m-1}^{k} \alpha_m \, p\left(\frac{y}{\theta_m}\right) \qquad Ecuación\ (1)$$

Donde α1, α2,…, αk son probabilidades mezclantes (probabilidades a priori) que nos indican el grado de importancia de cada uno de los k modos, cada Θm es el vector de parámetros que define la m-ésima componente, Θ = {Θ1,…, Θk, α1,…, αk} es el conjunto completo de parámetros necesarios para especificar la mixtura, por supuesto las αm deben satisfacer:

$$\alpha_m \geq para\ todo\ m = 1, \dots, k\ y \sum_{m-1}^{k} \alpha_m = 1 \quad Ecuación\ (2)$$

5.4.2. Algoritmos de agrupamiento no paramétricos

5.4.2.1. Algoritmos de agrupamiento jerárquico Los algoritmos de agrupamiento jerárquico tienen por objetivo agrupar clústeres con el objetivo de formar nuevos, igualmente separar uno para dar origen a otro grupo, de tal forma que, si se efectúa esta separación división) o una unión (aglomeración), se minimice alguna distancia o se maximice alguna medida de similitud. Los métodos jerárquicos se subdividen en aglomerativos y disociativos.

5.4.2.1.1. Métodos Aglomerativos. También conocidos como ascendentes, comienzan el análisis con tantos grupos como individuos haya. A partir de estas unidades iniciales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado.

Sea n el conjunto de individuos de la muestra, de donde resulta el nivel K = 0, con n grupos. En el siguiente nivel se agruparán aquellos dos individuos que tengan la mayor similitud (o menor distancia), resultando así n − 1 grupos; a continuación, y siguiendo con la misma estrategia, se agruparán en el nivel posterior, aquellos dos individuos (o clústeres ya formados) con menor

distancia o mayor similitud; de esta forma, en el nivel L tendremos n − L grupos formados. Si se continúa agrupando de esta forma, se llega al nivel L = n − 1 en el que solo hay un grupo, formado por todos los individuos de la muestra. Esta manera de formar nuevos grupos tiene la particularidad de que, si en un determinado nivel se agrupan dos clústeres, estos quedan ya jerárquicamente agrupados para el resto de los niveles. (“Métodos Jerárquicos de Análisis Cluster,” n.d.)

Dentro de los métodos aglomerativos jerárquicos tenemos:

- Estrategia de la distancia mínima o similitud máxima.
- Estrategia de la distancia máxima o similitud mínima.
- Estrategia de la distancia, o similitud, promedio no ponderada. (Weighted aritmethic average)
- Estrategia de la distancia, o similitud, promedio ponderada. (unweighted aritmethic average)
- Métodos basados en el centroide.
- Método de Ward.
- Métodos Disociativos.

También llamados descendentes, constituyen el proceso inverso al anterior. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

Se puede decir que la filosofía de los métodos Aglomerativos puede mantenerse para este otro tipo de procedimientos en lo que concierne a la forma de calcular la distancia entre los grupos, si bien, como es lógico, al partir de un grupo ´único que hay que subdividir, se seguirá la estrategia

de maximizar las distancias, o minimizar las similitudes, puesto que buscamos ahora los individuos menos similares para separarlos del resto del conglomerado.(“Métodos Jerárquicos de Análisis Cluster,” n.d.)

Esta clase de métodos son esencialmente de dos tipos:

- **Monotéticos**, los cuales dividen los datos sobre la base de un solo atributo y suelen emplearse cuando los datos son de tipo binario.
- **Politéticos,** cuyas divisiones se basan en los valores tomados por todas las variables.

5.4.2.2. Algoritmos de agrupamento no jerárquico o particional. Dado D, un conjunto de datos de n objetos, y k, el número de clústeres a formar, un algoritmo de partición organiza los objetos en k particiones ($k \leq n$), donde cada partición representa un clúster. Los clústeres se forman para optimizar un criterio de partición objetivo, como una función de disimilitud basada en la distancia, de modo que los objetos dentro de un clúster sean "similares", mientras que los objetos de diferentes clústeres sean "diferentes" en términos de los atributos del conjunto de datos. (Han et al., 2011)

Estos algoritmos asumen un conocimiento a priori del número de clusters en que debe ser dividido el conjunto de datos, llegan a una división en clases que optimiza un criterio predefinido o función objetivo. Entre los algoritmos que emplean esta técnica podemos mencionar: K-Means, k-medoids, métodos basados en densidad, algoritmo DENCLUE, algoritmo SNN y LDA.

5.4.2.3. Algoritmo K-means. Entre el grupo de algoritmos particionales es uno de los más conocidos y de los más simples, este divide la base de datos dada en k grupos de forma a priori, es decir, el algoritmo k-means toma el parámetro de entrada, k, y divide un conjunto de n objetos en k clústeres para que la similitud resultante del intra-clúster sea alta, pero la similitud entre clústeres sea baja. La similitud de proximidad se mide con respecto al valor medio de los objetos en un grupo, que se puede ver como centroide o centro de gravedad del grupo. (Han et al., 2011)

El algoritmo k-means funciona de la siguiente manera. En primer lugar, selecciona aleatoriamente k de los objetos, cada uno de los cuales representa inicialmente una media o centro del grupo. Para cada uno de los objetos restantes, un objeto se asigna al clúster al que es más similar, según la distancia entre el objeto y la media del clúster. Luego calcula la nueva media para cada grupo. Este proceso se repite hasta que la función de criterio converge. Normalmente, se utiliza el criterio de error cuadrado, definido como (Ver ecuación 3):

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \quad Ecuación\ (3)$$

Donde E es la suma del error cuadrado para todos los objetos en el conjunto de datos; p es el punto en el espacio que representa un objeto dado; y mi es la media del grupo Ci (tanto p como m_i

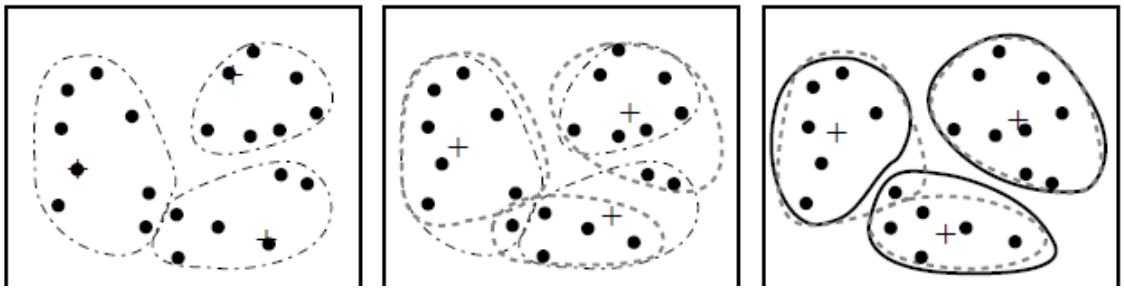

Figura 2. Agrupación de un conjunto de objetos basado en el método k-Means. Tomado de libro Data mining Han (2011)

Son multidimensionales). En otras palabras, para cada objeto en cada grupo, la distancia desde el objeto a su centro de grupo se cuadra, y las distancias se suman. Este criterio intenta hacer que los clusters k resultantes sean tan compactos y tan separados como sea posible.

El problema del empleo de estos esquemas es que fallan cuando los puntos de un grupo están muy cerca del centroide de otro grupo, también cuando los grupos tienen diferentes tamaños y formas. (Han et al., 2011)

5.4.2.4. Algoritmo K-medoids. El algoritmo k-means es sensible a valores atípicos porque un objeto con un valor extremadamente grande puede distorsionar sustancialmente la distribución de datos. Este efecto es particularmente exacerbado debido al uso de la función de error cuadrado (ver ecuación 1).

Para disminuir tal sensibilidad, En lugar de tomar el valor promedio de los objetos en un clúster como punto de referencia, podemos elegir objetos reales para representar los clústeres, utilizando un objeto representativo por clúster. Cada objeto restante se agrupa con el objeto representativo al que es más similar. El método de partición se realiza según el principio de minimizar la suma de las diferencias entre cada objeto y su punto de referencia correspondiente. Es decir, se usa un criterio de error absoluto, definido como (Ver ecuación 2):

$$E = \sum_{j=1}^{k} \sum_{p \in C_i} |p - O_j| \quad Ecuación\ (4)$$

Donde E es la suma del error absoluto para todos los objetos en el conjunto de datos; p es el punto en el espacio que representa un objeto dado en clúster *Cj*; y *oj* es el objeto representativo de *Cj*. En general, el algoritmo itera hasta que, eventualmente, cada objeto representativo es en

realidad el objeto medoide, o el objeto ubicado más centralmente, de su clúster. Esta es la base del método k-medoids para agrupar n objetos en k clústeres.

Dicho esto, la agrupación de k-medoids funciona así: los objetos representativos iniciales (o semillas) se eligen arbitrariamente. El proceso iterativo de reemplazar objetos representativos por objetos no representativos continúa mientras se mejore la calidad del agrupamiento resultante. Esta calidad se estima utilizando una función de costo que mide la disimilitud promedio entre un objeto y el objeto representativo de su clúster. (Han et al., 2011)

5.4.2.4.1. Algoritmo CURE. Este algoritmo es un hibrido entre los dos enfoques jerárquico y particional donde busca emplear las ventajas de ambos métodos y de eliminar las limitaciones; este en vez de utilizar un único punto (centroide) que represente el grupo se emplea un numero c de números representativos del grupo, así, la similitud entre dos grupos se mide por la similitud del par de puntos representativos más cercanos, uno de cada grupo.

Para tomar los puntos más representativos de cada grupo, se selecciona los c puntos más dispersos del grupo y los atrae al centroide bajo un factor de contracción α, en cada paso se unen los dos grupos más cercanos y una vez unido se vuelve a calcular para éste su centro y los c puntos representativos asociados al grupo. (Pascual, Pla, & Sánchez, n.d.-b).

5.4.2.5. Métodos basados en la densidad. Los métodos de partición jerárquico y no jerárquico son adecuados para encontrar agrupaciones de forma esférica o convexa. Dicho de otra manera, funcionan muy bien para grupos compactos y bien separados. Los algoritmos basados en densidad no necesitan definir previamente el número de clusters, y que se puede identificar grupos con cualquier forma contengan ruido y valores típicos. Estos algoritmos no necesitan de un numero k grupos a priori

5.4.2.5.1. Algoritmo DBSCAN. (Agrupamiento espacial basado en densidad de aplicaciones con ruido) es un algoritmo que hace crecer regiones con una densidad suficientemente alta en clústeres y descubre grupos de forma arbitraria en bases de datos espaciales con ruido, no es necesario determinar el número de clústers a crear.

El algoritmo define un clúster como un grupo de puntos conectados dependiendo de la densidad.

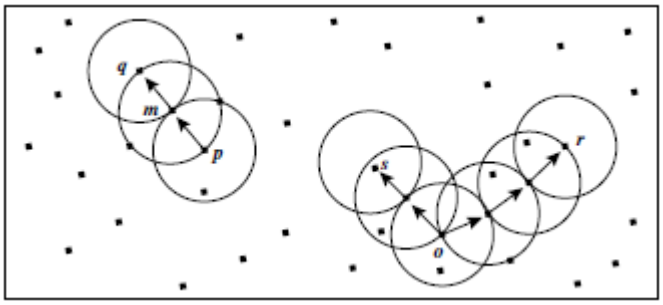

Figura 3. Accesibilidad y conectividad de densidad en clústeres basados en densidad Tomado de Han 2011

DBSCAN inicia seleccionando un punto P arbitrario, donde si P es un punto central, se comienza a construir un grupo, dentro de este grupo se ubican todos los objetos denso-alcanzables desde p, por el contrario, si P no es un punto central, se selecciona otro punto vecino del conjunto de datos; el proceso finaliza cuando todos los puntos son seleccionados o visitados, los puntos que

quedan fuera de los grupos formados se llaman puntos ruido, los puntos que no son ni ruido ni centrales se llaman puntos borde. (Han et al., 2011)

5.4.2.5.2. Algoritmo OPTICS. Ordena puntos para identificar la estructura de agrupamiento, Aunque DBSCAN puede agrupar objetos con parámetros de entrada como *e* y MinPts, aún deja al usuario la responsabilidad de seleccionar valores de parámetros que conduzcan al descubrimiento de clusters aceptables. En realidad, este es un problema asociado con muchos otros algoritmos de agrupamiento. Tales configuraciones de parámetros generalmente se establecen empíricamente y son difíciles de determinar, especialmente para conjuntos de datos de alta dimensión y del mundo real. Para ayudar a superar esta dificultad, se propuso un método de análisis de conglomerados llamado OPTICS, este, en lugar de producir una agrupación de conjuntos de datos explícitamente, calcula una ordenación de clúster para el análisis de clúster automático e interactivo. Este orden representa la estructura de agrupamiento basada en la densidad de los datos. Contiene información que es equivalente a la agrupación basada en la densidad obtenida a partir de una amplia gama de configuraciones de parámetros. El orden de clúster se puede utilizar para extraer información básica de clúster (como centros de clúster o clústeres de forma arbitraria) y también para proporcionar la estructura de clúster intrínseca. (Han et al., 2011)

Al examinar DBSCAN, podemos ver fácilmente que para un valor MinPts constante, los conglomerados basados en densidad con respecto a una mayor densidad (es decir, un valor menor para e) están completamente contenidos en los conjuntos conectados a densidad obtenidos respecto a una menor densidad. Por lo tanto, para producir un conjunto u orden de clústeres basados en densidad, podemos extender el algoritmo DBSCAN para procesar un conjunto de valores de parámetros de distancia al mismo tiempo. Para construir los diferentes agrupamientos

simultáneamente, los objetos deben procesarse en un orden específico. Esta orden selecciona un objeto que es accesible a la densidad con respecto al valor de e más bajo para que los conglomerados con mayor densidad terminen primero. (Han et al., 2011)

Según esta idea, se deben almacenar dos valores para cada objeto-distancia-núcleo y distancia de alcance:

- La distancia al núcleo de un objeto p es el valor Ɛ' más pequeño que hace que {p} sea un objeto central. Si p no es un objeto central, la distancia al núcleo de p no está definida.
- La distancia de alcance de un objeto q con respecto a otro objeto p es el mayor valor de la distancia del núcleo de p y la distancia euclidiana entre p y q.
- Si p no es un objeto central, la distancia de alcance entre p y q no está definida.

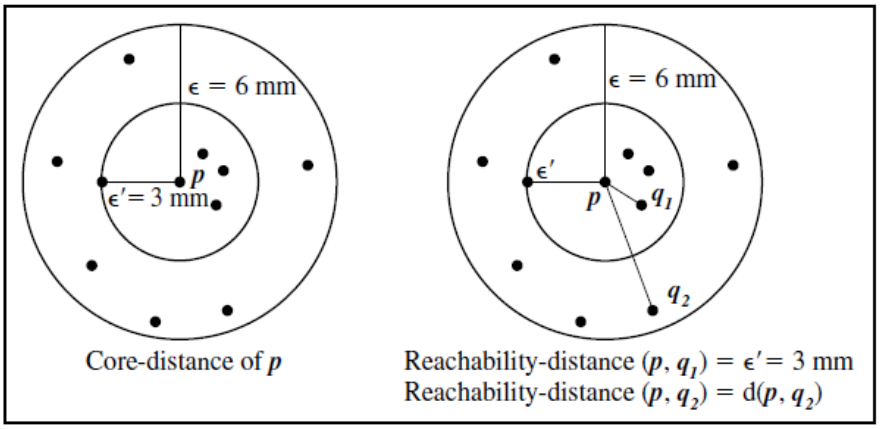

Figura 4. Terminología OPTICS. Tomado de Han 2011

*5.4.2.5.3. Algoritmo DENCLUE. "*Agrupación basada en funciones de distribución de densidad. DENCLUE es un método de agrupamiento basado en un conjunto de funciones de distribución de densidad. El método se basa en las siguientes ideas: (1) la influencia de cada punto de datos se puede modelar formalmente usando una función matemática, llamada función de influencia, que describe el impacto de un punto de datos dentro de su vecindario; (2) la densidad total del espacio de datos se puede modelar analíticamente como la suma de la función de influencia aplicada a todos los puntos de datos; y (3) los conglomerados pueden determinarse matemáticamente identificando los atractores de densidad (*density attractors)*, donde los atractores de densidad son máximos locales de la función de densidad global" (Han et al., 2011).

Sean (X, Y) objetos o puntos en F^d, un espacio de entrada d-dimensional. La función de influencia del objeto de datos 'y' en 'x' es una función $f^y B: F^d \rightarrow R_0^+$, que se define en términos de una función de influencia básica f_B:

$$f_B^X(x) = f_B(x, y) \qquad Ecuación\ (5)$$

Esto refleja el impacto de 'y' en 'x'. En principio, la función de influencia puede ser una función arbitraria que puede ser determinada por la distancia entre dos objetos en un vecindario; la función de distancia, d (x, y), debe ser reflexiva y simétrica. (Han et al., 2011).

5.4.2.5.4. Algoritmo SNN. Datos de alta dimensionalidad. Este algoritmo fue desarrollado debido a la existencia de bases de datos de alta dimensionalidad tales como textos y series de tiempo, así como la existencia de grupos de diferentes formas y tamaño.

Tiene su punto de partida en encontrar los datos más próximos de cada uno de los puntos a la base de datos y determina la similitud entre cada par de puntos en términos de vecinos más cercanos. Esto permite identificar puntos centrales, eliminar ruido y datos que no tienen ningún tipo de relación con los otros grupos o elementos de los grupos, constituye puntos alrededor de dichos puntos centrales, el concepto de similitud entre vecinos elimina asuntos que involucren diferencias en densidades, mientras que con el uso de puntos centrales se solucionan problemas referentes a la forma y el tamaño de las conglomeraciones. Para los pesos de los enlaces entre dos puntos en el grafo de los vecinos más cercanos compartidos (SNN) se toma en cuenta el ordenamiento de los vecinos más cercanos. Encuentra de manera natural la cantidad de grupos. (Pascual et al., n.d.-b)

5.4.2.5.5. Latent Dirichlet Allocation (LDA). "Es un modelo generativo que permite que conjuntos de observaciones puedan ser explicados por grupos no observados que explican por qué algunas partes de los datos son similares, es decir, se trata de un modelo probabilístico, ya que dice que un documento se crea mediante la selección de los temas y las palabras de acuerdo a las representaciones probabilísticas del texto natural" (Blei, Ng, & Jordan, n.d.; Rygielski et al., 2002).

La probabilidad inherente en los modelos de selección de cada palabra se deriva del hecho de que el lenguaje natural nos permite utilizar múltiples palabras diferentes para expresar la misma idea. Expresar esta idea en el modelo LDA, sirve para crear un documento sin un corpus, lo que se podría determinar cómo una distribución de temas. Para cada palabra del documento que se está generando, se escoge un tema de una distribución de Dirichlet de temas. A partir de ese tema, se coge una palabra elegida al azar basada en otra distribución de probabilidad condicionada en ese

tema. Esto se repite hasta que el documento se ha generado. (Benitez Andradez, José; Valvuena Lopéz, 2011)

En LDA, cada documento puede verse como una mezcla de varias categorías. Esto es similar a Probabilistic Latent Semantic Analysis (pLSA), excepto que en LDA se asume que la distribución de categorías tiene una distribución a priori de Dirichlet. La clave en LDA es que las palabras siguen una hipótesis de bolsa de palabras o, más bien que el orden no importa, que el uso de una palabra es ser parte de un tema y que comunica la misma información sin importar dónde se encuentra en el documento, este supuesto es necesario para que las probabilidades sean intercambiables y permitan una mayor aplicación de métodos matemáticos. Las palabras que aparecen con menos frecuencia en los documentos únicos, pero son comunes en muchos documentos diferentes probablemente son indicativo de que existe un tema común entre los documentos. Cuando se genera un resumen, la capacidad de recoger los matices de los temas del documento permiten que la información más relevante sea incluida con menos posibilidades de repetición y dar así un resumen mejor. (Benitez Andradez, José; Valvuena Lopéz, 2011).

5.5. Silhouette Method

El método de silueta (silhouette method) es utilizado para estudiar la distancia de separación entre los grupos resultantes después de agrupar una serie de datos. El gráfico de silueta muestra que tan cerca este cada punto de un grupo a los puntos en los grupos vecinos y, por lo tanto, proporciona una forma de evaluar visualmente parámetros como el número de grupos. Esta medida tiene un rango de [1, -1] (Carmona., 2015).

Los valores de silueta (como son denominados dichos valores) cerca de +1 indican que la muestra está muy lejos de los grupos vecinos. Un valor cero (0) indica que la muestra está muy

cerca del límite de decisión entre los grupos vecinos y los valores negativos indican que esas muestras se asignaron a un grupo incorrecto.

Para cada observación i, el silhouette coeficient (s_i) se obtiene de la siguiente manera:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad Ecuación\ (6)$$

Donde:

- a_i: promedio de las distancias entre la observación i y el resto de las observaciones.
- b_i: menor de las distancias promedio entre i y el resto de los clusters (la distancia al cluster más próximo)

5.6. Elbow Method.

Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de Clusters (desde 1 a N Clusters), siendo la inercia la suma de las distancias al cuadrado de cada objeto del Cluster a su centroide:

$$Inercia = \sum_{i=0}^{N} \|xi - u\|^2 \quad Ecuación\ (7)$$

Una vez obtenidos los valores de la inercia tras aplicar el K-means de 1 a N Clusters, se representa en una gráfica lineal la inercia respecto del número de Clusters. En esta gráfica se debería de apreciar un cambio brusco en la evolución de la inercia, teniendo la línea representada una forma similar a la de un brazo y su codo. El punto en el que se observa ese cambio brusco en la inercia nos dirá el número óptimo de Clusters a seleccionar para esa data set; o, dicho de otra

manera: el punto que representaría al codo del brazo será el número óptimo de Clusters para ese conjunto de datos.

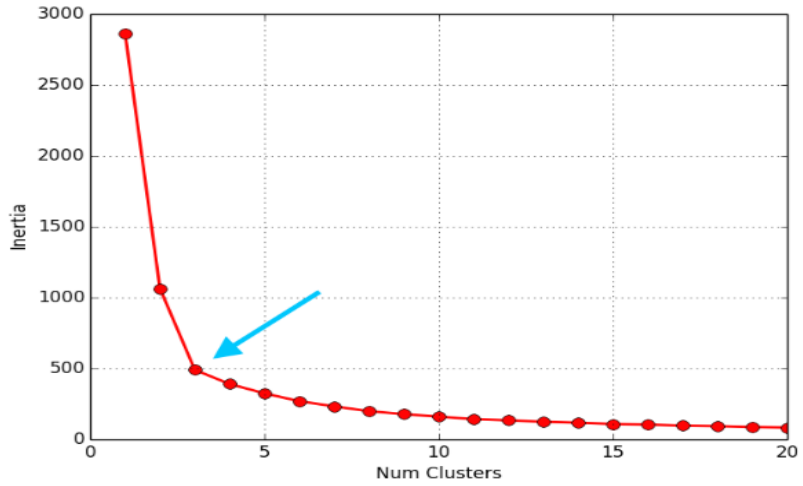

Figura 5. Cantidad optima de clúster. Obtenido de: Ricardo Moya, (2016)

5.7. Minería de textos en Twitter con R

El paquete TwitteR de R permite obtener los datos que cede Twitter a través de sus APIs. Con este paquete se puede ver los tweets de las cuentas o usuarios de interés, el número de seguidores, a las personas que siguen, sus timelines, la cantidad de reacciones o re-tweets que tienen, identificar los tweets relacionados con algunas temáticas, la localización del usuario que publica, etc. (Isabel & Moreno, 2017). El paquete de R tiene diversidad de usos en aspectos que involucran redes sociales y análisis de datos, entre ellos se encuentran:

- Minería de textos. Los tweets permiten identificar tendencias o temas de interés dentro de la comunidad.
- Dependiendo de la localización geográfica se puede determinar, a partir de TwitteR, identificar temas que son tendencia y la periodicidad con la que se habla de dichos temas.

- Debido a los rangos temporales identificar tendencias teniendo en cuenta las palabras que emplean los usuarios.

6. Base de datos

Las redes sociales hoy por hoy constituyen un fenómeno social, político y económico que permite compartir, comunicar, interactuar y conocer el mundo a través de herramientas electrónicas. Twitter aparece como una posibilidad de microblogging que permite, a tevés de un texto de 140 caracteres, expresar opiniones o posiciones sobre temas que son tendencia (Fainholc, 2011).

Twitter cuenta con 339 millones de usuarios activos para enero de 2020, a pesar de no tener el crecimiento acelerado como otras redes sociales (Facebook e Instagram) su fortaleza radica en la disponibilidad de información en tiempo real, es para muchas marcas el medio de comunicación oficial y permite a las marcas hacer campañas de carácter social a través de esta plataforma. Por lo anterior Twitter se convierte en una estrategia de marketing digital.

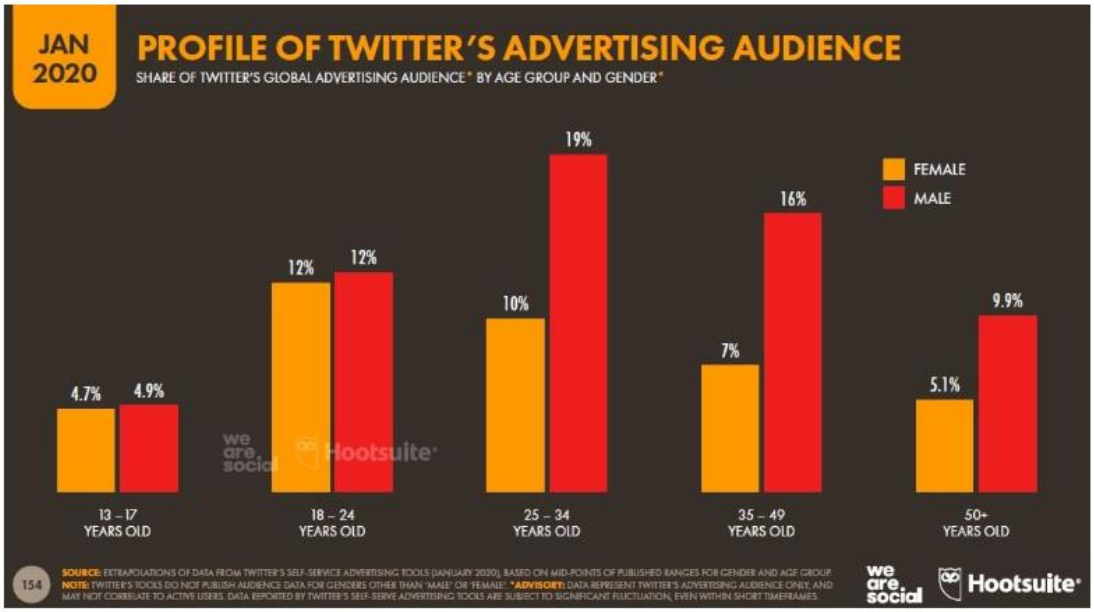

Figura 6. Perfil de audiencia publicitaria en Twitter. Obtenido de: marketing Digital (Mejía, 2020)

Usando la minería de texto se hace un análisis descriptivo de una base de datos obtenida de Twitter, que a través de su API permite indagar sobre el origen de los datos para esta investigación. Para la consolidación de dicha base se usó la cuenta oficial en Twitter de la Universidad Industrial de Santander (@UIS), se monitoreó dicha cuenta a lo largo del mes de marzo y se pudo recopilar 3217 tweets que comprenden desde el mes de octubre de 2019 hasta el mes de marzo de 2020.

Para el procesamiento y limpieza de la base de datos se hizo lo siguiente:

- Creación de un Corpus que contiene los Tweets a analizar a partir de la base de datos que se consolidó.
- Limpieza de símbolos, enlaces web y palabras vacías (stopwords) en el contenido del corpus.

- Creación de una matriz término-documento que permite describir la frecuencia de los términos que se producen en una colección de documentos.

Las palabras más frecuentes en el contenido de los Tweets fueron las siguientes (representación en un gráfico de barras):

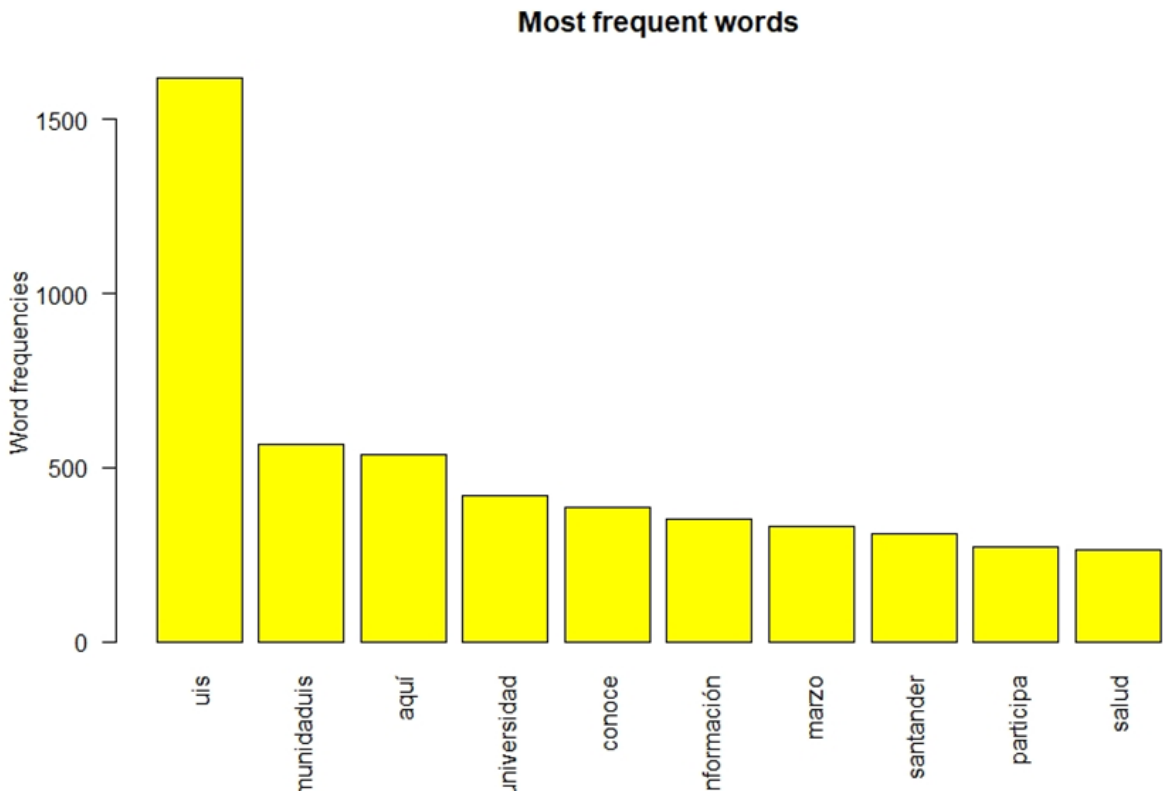

Figura 7. Palabras más frecuentes

- Nube de palabras con las palabras más repetidas en los textos.

Figura 8. Nube de palabras

- Se eliminan los Tweets que no tienen palabras.

7. Evaluación de los métodos de agrupamiento DBSCAN y LDA

7.1. Algoritmo DBSCAN

7.1.1. Cálculo de parámetros óptimos para DBSCAN. Se realizó la selección de 20 elementos para MinPts (mínimo de puntos para formar un cluster), para mejorar el análisis este número de elementos se determinó a través de la realización de pruebas en dónde se encontró que para valores inferiores de este parámetro la cantidad de agrupamientos obtenidos era muy alta lo que dificultaba el análisis de datos, la obtención de la información y la comparación con el algoritmo LDA.

Para la obtención de EPS se utiliza la función KNNdisplot del algoritmo DBSCAN la cual sirve para realizar un cálculo rápido de las distancias K-vecinas en la matriz de puntos siendo la distancia

óptima (EPS) la que represente el punto donde se evidencie un cambio significativo en la pendiente de la curva como se evidencia en la Figura 16.

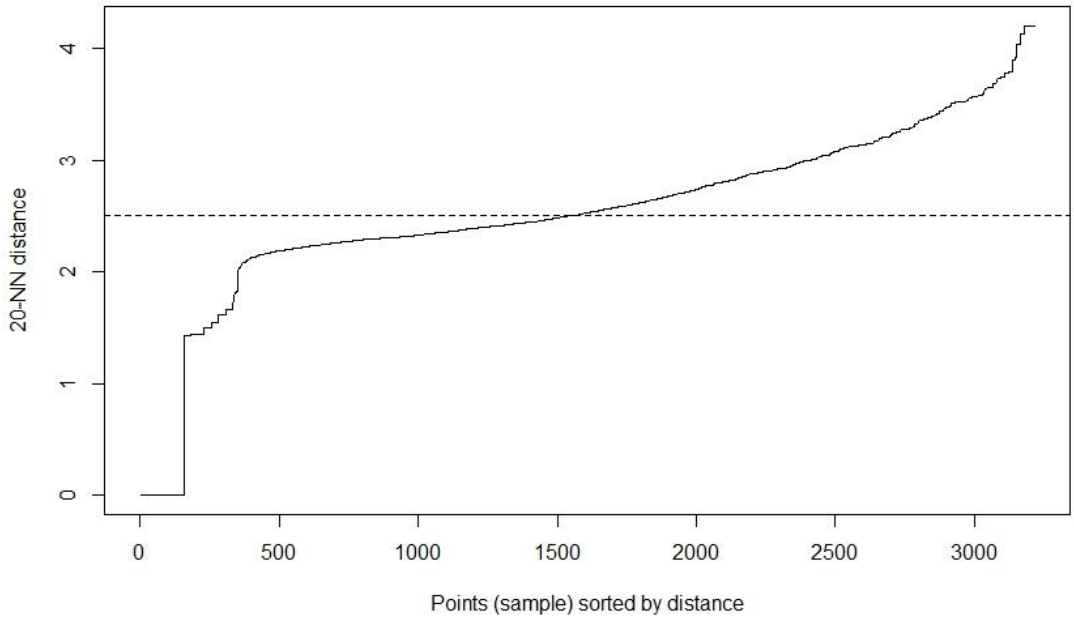

Figura 9. Gráfico EPS

Se escoge un valor de EPS igual a 3.0.

Con la ejecución del algoritmo DBSCAN con los parámetros menciones se obtiene que nueve clusters, además de 216 elementos considerados ruido o valores extremos (outliers).

Tabla 3.

Distribución de elementos para cada cluster.

```
Cluster.dbscan
```

0	1	2	3	4	5	6	7	8	9
216	2599	78	26	148	23	49	20	21	36

En la tabla 2 el número de elementos pertenecientes a cada uno de los clusters formados.

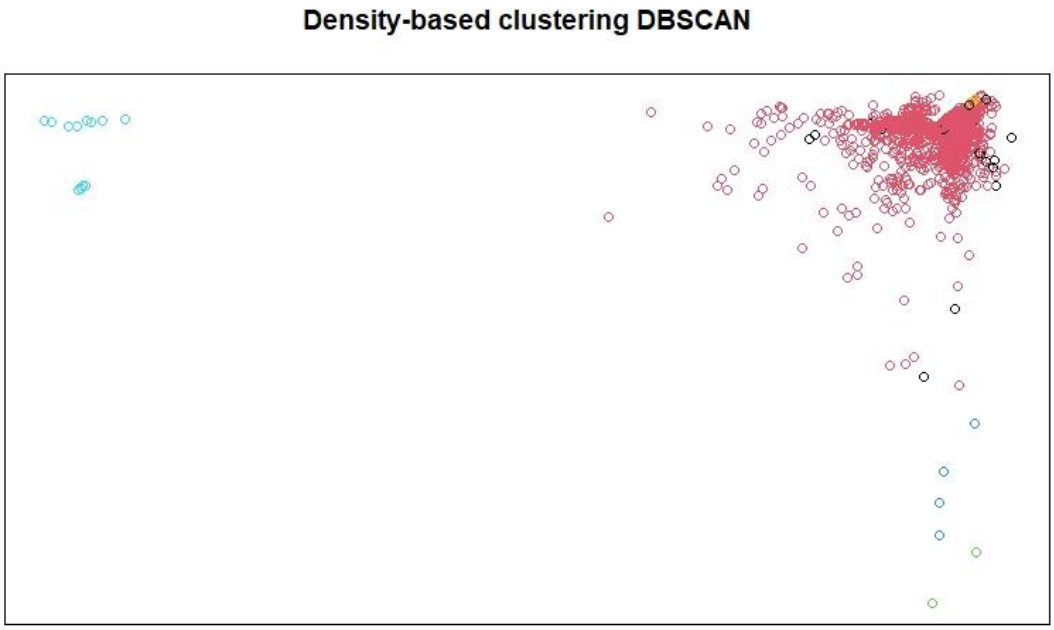

Figura 10. Gráfico de dispersión del algoritmo DBSCAN

De acuerdo al gráfico de dispersión del algoritmo DBSCAN se observa que los tópicos formados no se encuentran bien definidos gráficamente, esto se debe a que el 80% de los elementos analizados pertenecen a un solo agrupamiento. Este resultado era esperado, ya que la palabra UIS es la de uso más frecuente en los Tweets estudiados y aparece en la mayoría de los clusters estudiados.

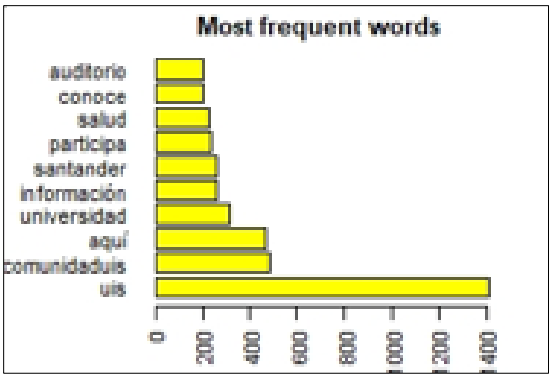

Figura 11. Frecuencia de palabras cluster N°1 de DBSCAN

Para el cluster N°1 se puede deducir, basado en las palabras con más frecuencia, tales como: auditorio, participa y comunidad UIS que el tema central en este período de tiempo es acerca de los eventos culturales realizados en las instalaciones de la universidad y la invitación a toda la comunidad UIS a participar de estas actividades. También es importante resaltar que estas palabras, debido a su alta frecuencia de uso, aparecen en la mayoría de agrupaciones que se hicieron para el algoritmo DBSCAN.

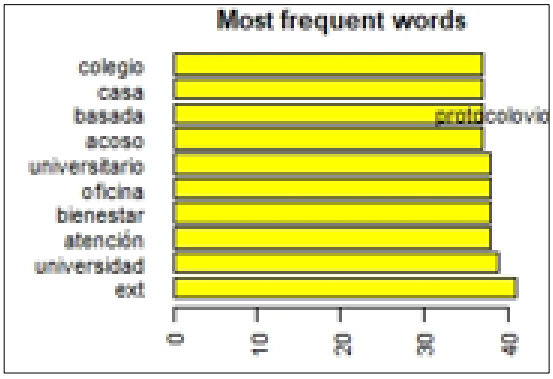

Figura 12. Frecuencia de palabras cluster N°2 de DBSCAN

Para el segundo cluster uno de los términos más frecuentes es EXT que hace referencia a la extensión telefónica de bienestar universitario y a los diferentes programas que esta brinda la universidad concernientes a situaciones de acoso de género.

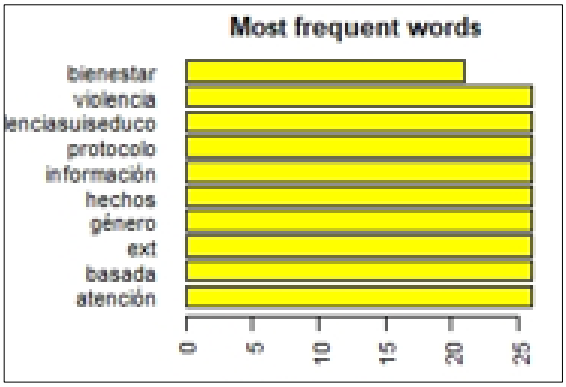

Figura 13. Frecuencia de palabras cluster N°3 de DBSCAN

En la tercera agrupación también hace referencia a la extensión de bienestar universitario, pero enfocado al tema de violencia al interior de la universidad y casos de violencia de género, esto debido a los acontecimientos que sucedieron en relación a dos estudiantes de la universidad que fueron víctimas de un feminicidio en los alrededores del campus universitario.

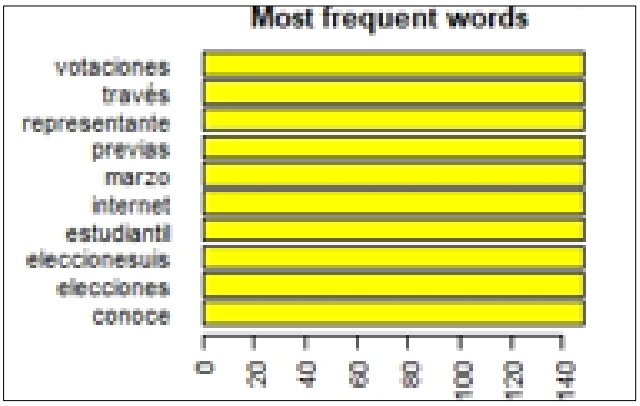

Figura 14. Frecuencia de palabras cluster N°4 de DBSCAN

En el cluster N°4 se evidencia la invitación a participar en las elecciones de los representares estudiantiles las cuales tuvieron lugar en el mes de marzo. Además, se les invitaba a realizar las votaciones a través de internet.

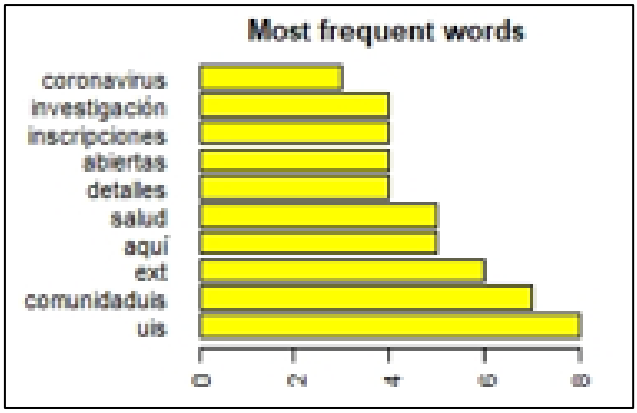

Figura 15. Frecuencia de palabras cluster N°5 de DBSCAN

En el tópico N°5 se empieza a evidenciar las opiniones sobre temas relacionados con la salud y el llamado a la prevención para evitar la propagación del Coronavirus. Además, se abren inscripciones para los programas de investigación y extensión de la UIS para el primer semestre del 2020.

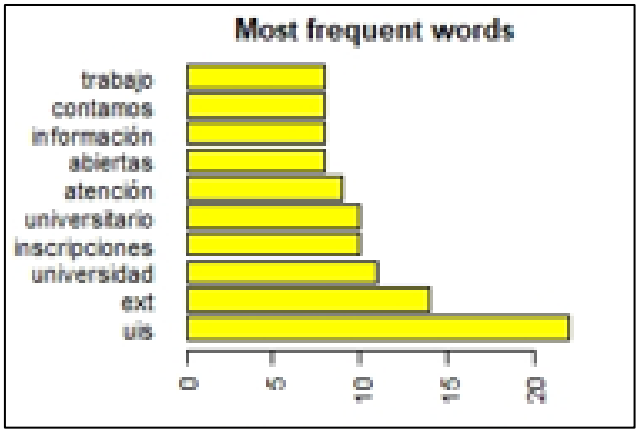

Figura 16. Frecuencia de palabras cluster N°6 de DBSCAN

Al igual que en cluster N°5, en la agrupación N°6 se ve que se sigue en el proceso de inscripción para los programas brindados por la UIS, pero se enfoca a temas relacionados con la información y atención a la comunidad.

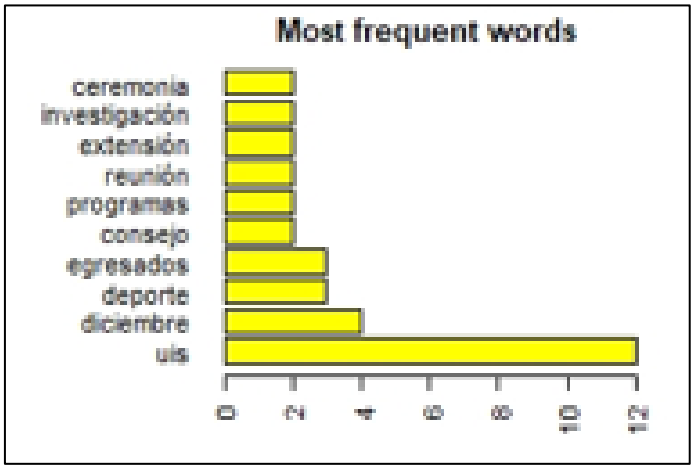

Figura 17. Frecuencia de palabras cluster N°7 de DBSCAN

Para el cluster N°7 se ve que hubo una reunión de los egresados de los programas académicos, también hace referencia a la ceremonia de grados llevada a cabo en el mes de diciembre del 2019.

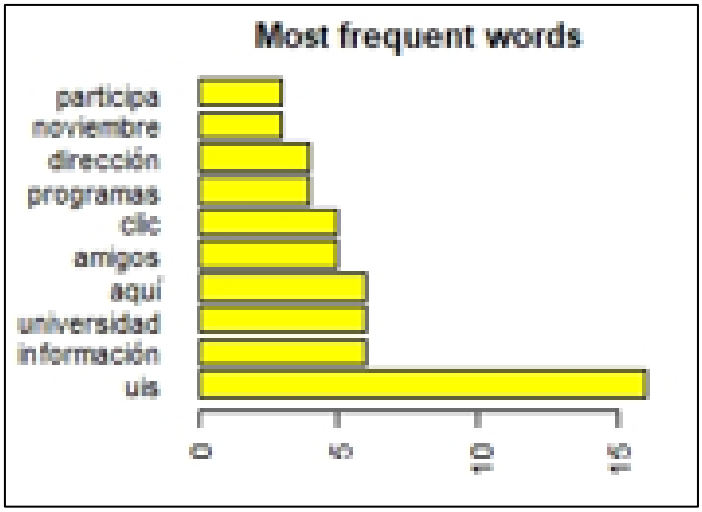

Figura 18. Frecuencia de palabras cluster N°8 de DBSCAN

En el N°8 se invita a participar en el mes de noviembre en las actividades de vecinos y amigos, cuya información se puede encontrar en la cuenta oficial de Twitter de la UIS.

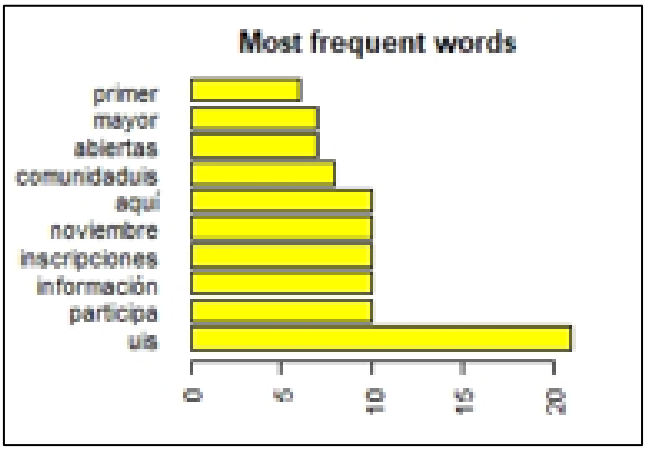

Figura 19. Frecuencia de palabras cluster N°9 de DBSCAN

Para el cluster N°9 se invita a la comunidad a participar de las inscripciones que se abrieron en el mes de noviembre de 2019.

De acuerdo a la frecuencia de palabras para cada cluster representadas de la figura 13 hasta la figura 21 se evidencia que la palabra UIS se encuentra en seis de los nuevos tópicos formados.

7.2. Algoritmo LDA

Antes de aplicar el algoritmo de agrupamiento LDA, se realizó una prueba para determinar la cantidad óptima de Clusters a hacer con los datos obtenidos después de la limpieza. Para futuras referencias se calculó el número de clusters para el algoritmo K-means. Para ello se utilizó dos métodos:

- Método del codo (Elbow Method)

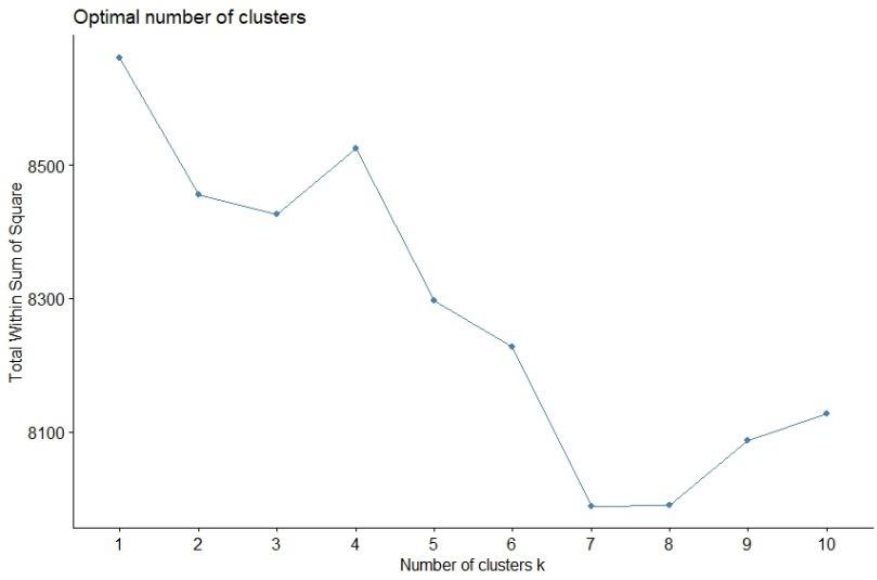

Figura 20. Elbow Method

En base a la figura 14 se puede evidenciar que existen tres puntos (2, 5 y 7) donde el cambio de la inercia de los datos es brusco en comparación a la línea de tendencia de la gráfica, los cuales representan el número óptimo de cluster. Sin embargo, presenta cierto tipo de confusión en su análisis pues no se ve una tendencia definida.

- Método de la Silueta (Silhouette Method)

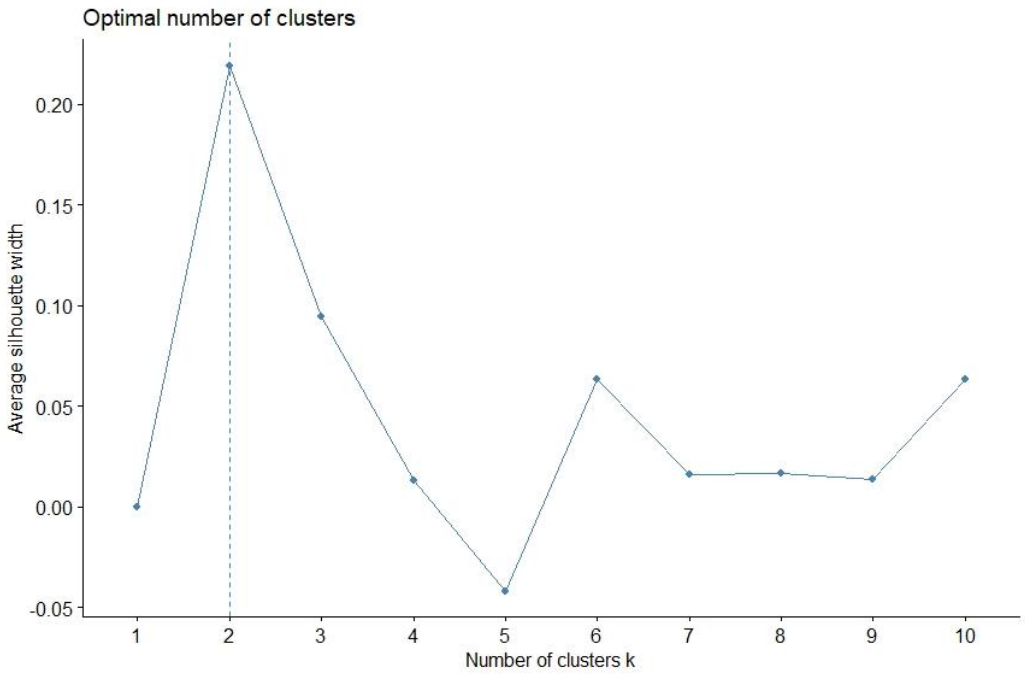

Figura 21. Silhouette Method

Como se ve en la Figura 15 el número los puntos más altos del coeficiente de silueta están en los clusters dos, seis y diez. Como se ve en el gráfico los valores del coeficiente de silueta son muy bajos en comparación a los resultados esperados, acercándose al cero en la mayoría de los casos lo que significa que no hay separación clara de los clusters, también se evidencia valores por debajo de cero (cinco clusters) lo que representa que para una cantidad de cinco clusters la agrupación de datos no va a ser la adecuada y sus elementos no van a tener relación entre sí.

Cabe mencionar que estos dos métodos (codo y silueta) se aplican sólo para el algoritmo LDA en este caso, dado que DBSCAN determina por si sólo el número óptimo de tópicos a formar dependiendo de la cantidad de datos que se tengan para analizar y los valores de los parámetros iniciales seleccionados (EPS y MinPts).

7.2.1. Aplicación LDA. Para este algoritmo de agrupamiento se necesita determinar la cantidad optima de tópicos que vamos a obtener, para esto se utilizó como referencia el método del codo y la silueta anteriormente mencionados, así como la aplicación de la función **FindTopicsNumber** comparando para ello 4 métricas: Griffiths2004, CaoJuan2009, Arun2010 y Deveaud2014.

Se escogió un número máximo de tópicos de 10 para la simplificación del análisis.

El objetivo es minimizar las métricas CaoJuan2009 y Arun2010 y maximizar las métricas Griffiths2004 y Deveaud2014.

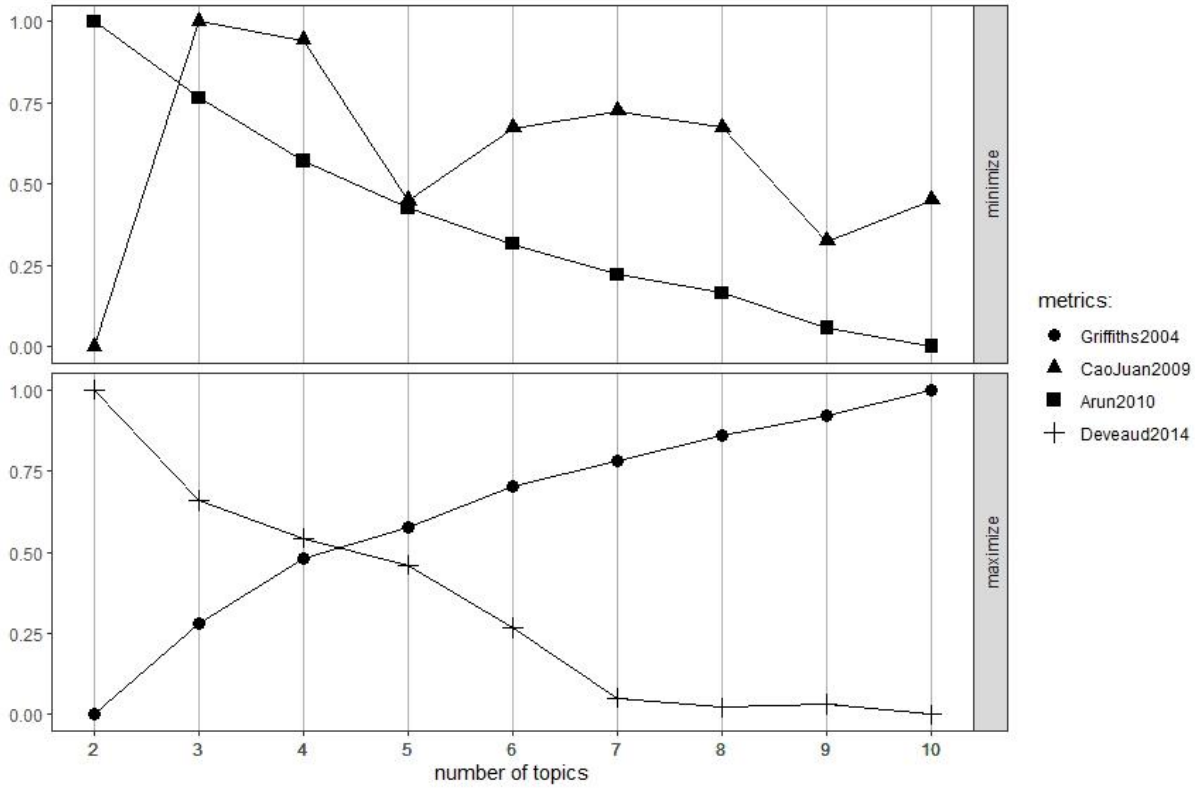

Figura 22. Métricas para determinar número de clusters óptimo LDA

Como se observa en la figura 22 las métricas CaoJuan2009 y Arun2010 coinciden en los puntos cinco y nueve. Mientras que las métricas Deveaud2014 y Griffits2004 no coinciden, pero los valores de Griffits2004 si aumentan en estos puntos. Se concluye que para LDA basadas en estas métricas el número óptimo de clusters es de cinco (k=5).

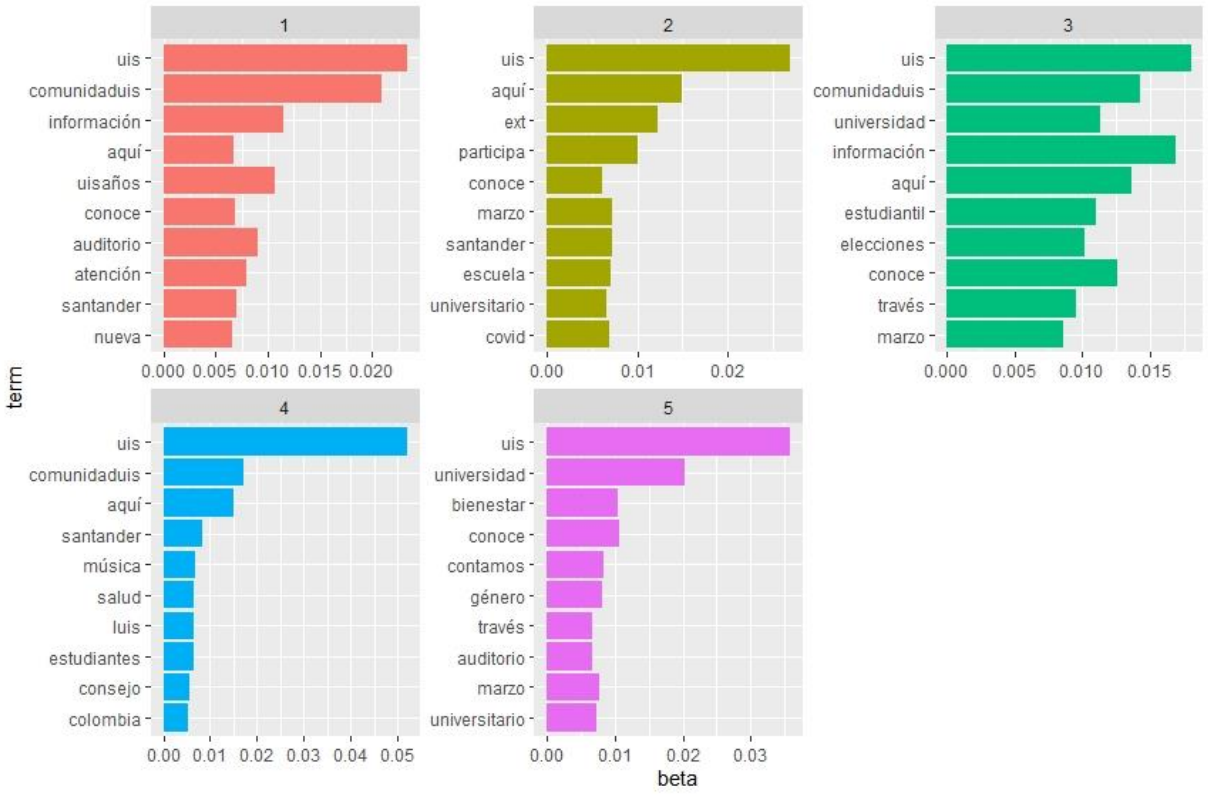

Figura 23. Cluster LDA para K=5

Para el LDA en términos generales se sigue evidenciando que se agrupan los mismos temas recurrentes que también encontró el DBSCAN referentes al departamento de extensión de la universidad, las elecciones del representante estudiantil, los programas de bienestar universitario enfocados a la atención de casos de violencia de género y los temas relacionados a la llegada del COVID-2019. Esto se evidencia en la figura 23, ya que los términos relacionados a estos temas son los de mayor probabilidad de aparición en los Tweets recopilaros durante este período de observación.

Para implicaciones del caso de estudio se decidió ejecutar el LDA con 9 clúster, debido a que este fue el número de agrupamientos que se realizaron en el DBSCAN esto con el fin de facilitar el análisis de los dos algoritmos bajo similares condiciones.

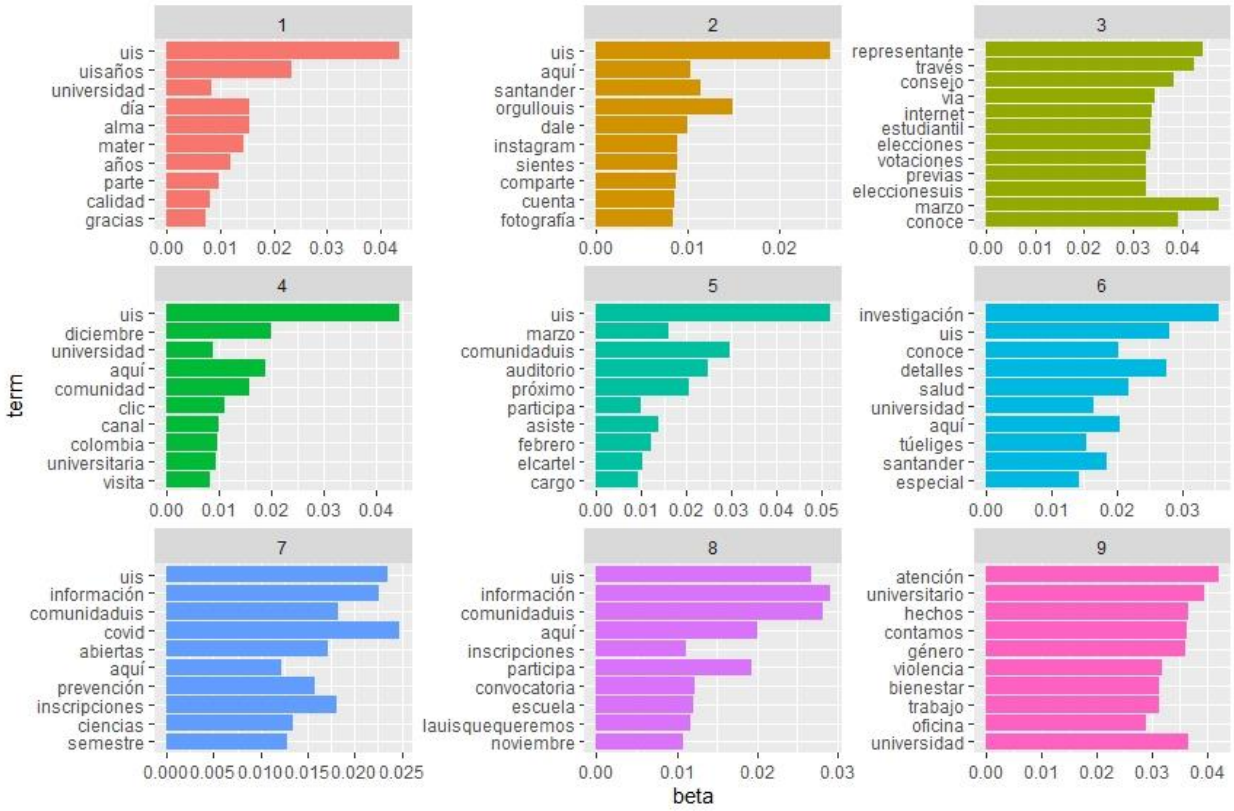

Figura 24. Clusters LDA para K=9

Con base a la figura 24 se puede observar que los elementos agrupados hacen referencia en general a los mismos temas agrupados por DBSCAN, sin embargo, es notable que LDA agrupa algunas palabras que para DBSCAN son consideradas ruido tales como Instagram, fotografía y comparte. Esto puede ser causado por la naturaleza misma del algoritmo ya que a diferencia de DBSCAN, LDA agrupa palabras en base a la probabilidad de que estas aparezcan en los documentos analizados.

7.3. Evaluación de los modelos.

Para la evaluación de los modelos se tuvo en cuenta un coeficiente llamado Shilohoutte coefficient, este coeficiente nos indica el nivel de coherencia que existe entre los elementos pertenecientes a un grupo. Sus valores se encuentran en un intervalo de [-1, 1], donde un valor

cercano a 1 indica que dicho elemento está bien relacionado con los demás elementos del mismo grupo y no tiene relación con los elementos de los demás grupos, valores cercanos a -1 prueban que el elemento está mal agrupado y los valores cercanos a 0 representan que la observación está entre dos cluster.

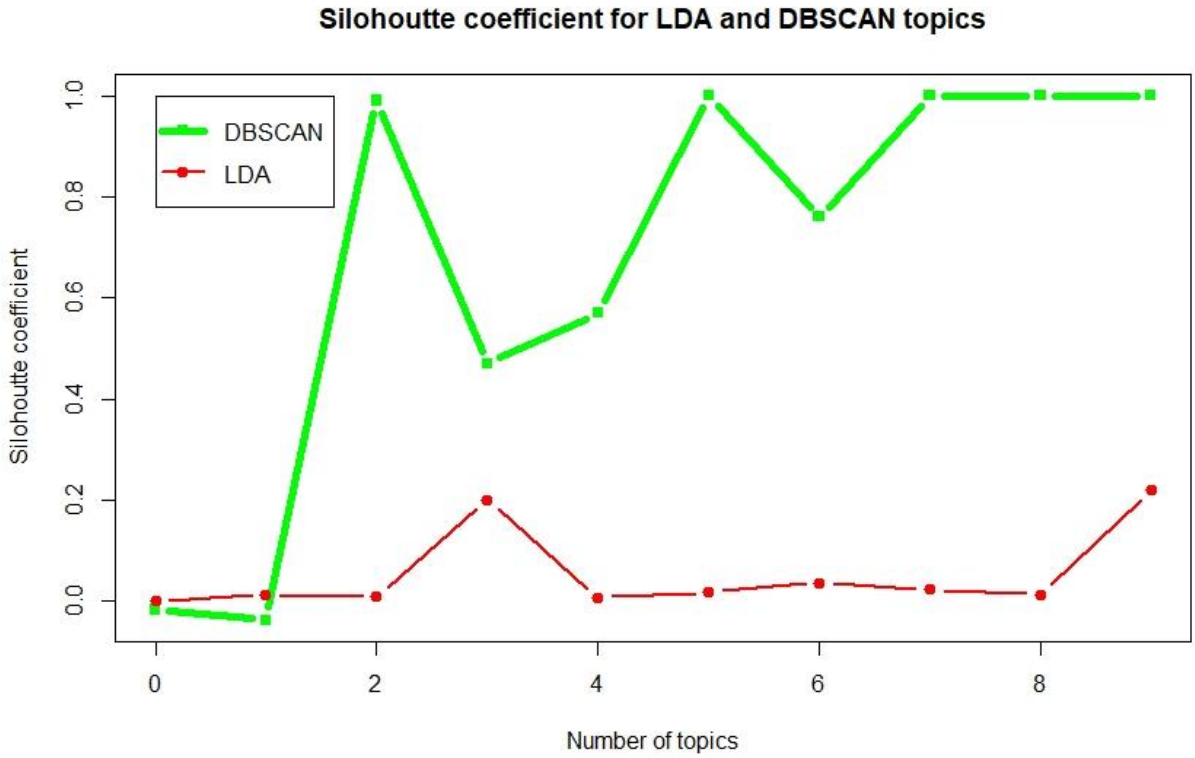

Figura 25. Coeficiente de silueta para los tópicos LDA y DBSCAN

En la Figura 25 se observa que los agrupamientos realizados por el algoritmo DBSCAN tienen un coeficiente de silueta más alto en comparación con los agrupamientos hechos por LDA, lo que indica que los tópicos del algoritmo DBSCAN tienen una mayor coherencia. También se observa que para el cluster N°1 de DBSCAN el coeficiente de silueta es cercano cero, evidenciando que algunos de sus elementos se encuentran en medio de varios grupos. En el caso de LDA, su

desempeño validado por el coeficiente de silueta en el caso de nueve clusters no fue el adecuado, ya que la mayoría de sus valores fueron muy cercanos a cero.

Para el cluster N°0 de DBSCAN representa todos los valores considerados como ruido para este algoritmo, mientras que el agrupamiento N°0 no existe para LDA.

8. Conclusiones

En este proyecto de investigación se consolidó una base de datos a partir de la API de Twitter compilando información de la cuenta oficial de la Universidad Industrial de Santander logrando recopilar 3217 tweets en un periodo de aproximadamente cinco meses. El objetivo principal era analizar el comportamiento de los algoritmos LDA (Latent Dirichlet Allocation) y DBSCAN en el agrupamiento de los datos obtenidos luego del procesamiento de la información obtenida para su posterior análisis y validación.

Para este caso de estudio se implementaron los métodos del codo y el coeficiente de silueta con, así como la evaluación de las métricas CaoJuan2009, Arun2010, Griffiths2004 y Deveaud2014 con el fin de determinar el número óptimo de clusters a formar. Estos métodos son aplicables al algoritmo LDA, ya que es necesario indicar el número de tópicos que requiere agrupar el algoritmo para su ejecución (en este caso el k óptimo fue de cinco grupos). Para fines de comparación con el otro modelo de agrupamiento se decidió realizar una prueba para nueve clusters dado que este fue el número de grupos que formó el algoritmo DBSCAN. Los parámetros iniciales para la ejecución del algoritmo DBSCAN (EPS y MinPts) fueron calculados mediante la función KNNdiplots encontrando valores para estos de 3.0 y 20 respectivamente.

El agrupamiento obtenido con la ejecución de los algoritmos seleccionados permitió la extracción de los principales temas que fueron tendencia en el periodo de tiempo analizado los cuales son de interés para la comunidad universitaria. Al hacer un análisis de los elementos de cada agrupamiento se evidencia una alta relación entre las palabras más frecuentes de cada grupo lo que demuestra un adecuado desempeño de los modelos.

De acuerdo a los resultados obtenidos con la implementación de los algoritmos mencionados en el caso de estudio y a la validación de los mismos, se llega a la conclusión que los agrupamientos realizados por DBSCAN fueron más coherentes que los obtenidos por LDA, dando como resultado que DBSCAN tiene un mejor desempeño con la base de datos estudiada.

Con lo anteriormente mencionado se evidencia que se dio cumplimiento a los objetivos principales del proyecto dejando espacio para mejoras y posibles aplicaciones en futuras investigaciones.

9. Recomendaciones

Para futuras investigaciones se recomienda realizar un exhaustivo preprocesamiento de los datos enfocado principalmente a la limpieza de los mismos dado que mucha información se pierde debido a problemas en la calidad del texto analizado, ocasionado por errores ortográficos o uso de jergas muy particulares que dificultan su análisis. Lo anteriormente mencionado es con el fin de enfocar los modelos de agrupamiento al estudio de la información relevante.

En el caso de este proyecto se encontraron dificultades con el cálculo de algunos parámetros iniciales de los algoritmos, especialmente el DBSCAN para lo cual se recomienda implementar

una variación más robusta como lo es el HDBSCAN, el cual requiere menos parámetros para su ejecución. Sin embargo, el DBSCAN mostró un mejor desempeño en comparación a otros modelos de agrupamiento tales como K-Means.

Otro aspecto importante a considerar es la selección de una adecuada base de datos, ya que se constató que varios de los datos analizados eran muy recurrentes y no había una gran variedad de temas, lo que ocasionó que la mayoría de los datos fueran agrupados en un solo tópico sesgando la investigación a un número muy reducido de temas obtenidos.

Referencias bibliográficas

Alsuwaidan, L., & Ykhlef, M. (2018). Interest-Based Clustering Approach for Social Networks. *Arabian Journal for Science and Engineering*, *43*(2), 935–947. Obtenido de: https://doi.org/10.1007/s13369-017-2800-z

Altamiranda, L., Peña, A., Ospino, M., Volpe, I., Ortega, D., & Cantillo, E. (2013). *Minería de datos como herramienta para el desarrollo de estrategias de mercadeo B2B en sectores productivos , afines a los colombianos : una revisión de casos*. 11.

Aluja, T. (2001). Data mining, between statistics and artificial intelligence. *Questiió*, *25*(3), 479–498.

Araujo, T., & Kollat, J. (2018). Communicating effectively about CSR on Twitter: The power of engaging strategies and storytelling elements. *Emerald Insight*.

Beltrán Martínez, B. (2014). Minería de datos. *Benemérita Universidad Autónoma de Puebla*, *1*(5), 67.

Benavides Rodríguez, C. (2016). *Análisis del uso de redes sociales en desastres*. Universidad de Oviedo.

Benitez Andradez, José; Valvuena Lopéz, J. (2011). Motores de Búsqueda Web: Estado del arte en Probabilistic Latent Semantic Analysis y en Latent Dirichlet Allocation aplicado a problemas de acceso a la información en la Web.

Berkhin, P. (2006). Agrupación de datos multidimensionales. Springer-Verlag Berlin Heidelberg.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). *Latent Dirichlet Allocation*. Journal of Machine Learning Research

Canzanello; del Canal; Rossmann. (2018). The future and social impct of big data analytics in Supply Chain Management: results from a Delphi Study. *Thecnological forecasting and*

social change, 135-136.

Carmona., S. (2015). *Data mining - Introducción al clustering en bioinformática.*

Cesur, R., Ceyhan, E. B., Kermen, A., & Sagiroglu, S. (2017). Determination of Potential Criminals in Social Network. *Journal od Science*.

Chu, K., Unger, J., Allem, J., Pattarroyo, M., Soto, D., Cruz, T., . . . Yang, C. (2015). Diffusion of Messages from an Electronic Cigarette Brand to Potential Users through Twitter. *Journal Citation Reports*.

Coumo, Salvatore; Maiorano, Francesco. (2018). Social Network data analysis and mining applications for the internet Data. *Special Issue Paper*.

Ding, Y., Yan, S., Zhang, Y., Dai, W., & Dong, L. (2016). Predicting the attributes of social network users using a graph-based machine learning method. *Computer Communications*, *73*, 3–11. Obtenido de: https://doi.org/10.1016/j.comcom.2015.07.007

Echeverri, L., Pena, A. M., Ospino, M. R., Volpe, I., Ortega, D., & Cantillo, E. (2014). Minería de datos como herramienta para el desarrollo de estrategias de mercadeo B2B en sectores productivos , afines colombianos: una revisión de casos . *Sotavento MBA*, 11.

Efrain Alberto Oviedo;Ana Isabel Oviedo; Gloria Liliana Vélez. (2015). Minería de datos: aportes y tendencias en el servicio de salud de ciudades inteligentes. *Revista Politécnica*, 116-117.

Fainholc, B. (2011). Un análisis contemporáneo del Twitter. *Revista Educación a Distancia*, 12.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal Of Statistical Software*, *25*(5), 1–54. https://doi.org/citeulike-article-id:2842334

Fernandez, Óscar. (1991). *Análisis de Cluster: interpretación y validación.* Revista de Sociología. Vol 37.

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. In M. Kamber (Ed.), *Elsevier* (Second, Vol. 12). Obtenido de: https://doi.org/10.1007/978-3-642-19721-5

HK, Chan; E, Lacka; R, Yee; MK, Lim. (2017). The role of social media data in operations and production management. INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH, 1-5

Isabel, A., & Moreno, V. (2017). *Técnicas estadísticas en Minería de Textos*. Universidad de Sevilla.

Mar, R. (2008). *Inteligencia Artificial. Ténicas, métodos y aplicaciones*. MC Graw Hill Interamericana.

Marcos, U. N. M. S. (2004). *Data Mining y el descubrimiento del conocimiento*. *7*, 83–86.

Mejía, L. J. (Junio de 2020). *Juan Carlos Mejía Llano*. Obtenido de https://www.juancmejia.com/marketing-digital/estadisticas-de-redes-sociales-usuarios-de-facebook-instagram-linkedin-twitter-whatsapp-y-otros-infografia/#:~:text=Twitter%20cuenta%20con%20m%C3%A1s%20339%20millones%20de%20usuarios%20activos%20en%20un%20mes.

Métodos Jerárquicos de Análisis Cluster. (n.d.).

Neme-chaves, S. R. (2018). *Twitter mining of the Juan Valdez coffee brand*. (May). https://doi.org/10.13140/RG.2.2.12863.41128

Ogan, C., & Varol, O. (2017). What is gained and what is left to be done when content analysis is added to network analysis in the study of a social movement: Twitter use during Gezi Park. *INFORMATION COMMUNICATION & SOCIETY*.

Pascual, D., Pla, F., & Sánchez, S. (n.d.-b). Algoritmos de agrupamiento. *Métodos Informáticos Avanzados*, 163–175.

Rashid, J., Muhammad, S., Shah, A., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. *Information Processing and Management*, *56*(6), 102060. Obtenido de: https://doi.org/10.1016/j.ipm.2019.102060

Rygielski, C., Wang, J.-C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society*, *24*(4), 483–502. Obtenido de: dehttps://doi.org/10.1016/S0160-791X(02)00038-6

Sadat, H., & Ali, C. (2017). *Toward a novel art inspired incremental community mining algorithm in dynamic social network*. 409–426. Obtenido de: https://doi.org/10.1007/s10489-016-0838-3

Tran, T., Drab, K., & Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Elsevier*.

Xia Liu; Alvin C. Burns; Yingjian Hou. (2017). An Investigation of brand-related user-generated content on Twitter. *Journal of Advertising*, 1-13.

Zhang, L., Dong, W., & Mu, X. (2018). Analysing the features of negative sentiment tweets. *Emerald Insight*.